Transactions of the ASME

TRANSACTIONS OF THE AMERICAN SOCIETY OF MECHANICAL ENGINEERS

VOLUME 79 APRIL 1957 NUMBER 3

Transactions

of The American Society of Mechanical Engineers

Published on the tenth of every month, except March, June, September, and December

Published monthly by The American Society of Mechanical Engineers. Publication office at 20th and Northampton Streets, Easton, Pa. The editorial department is located at the headquarters of the Society, 29 West Thirty-Ninth Street, New York 18, N. Y. Cable address, "Mechaneer," New York. Price $1.50 a copy, $12.00 annually for Transactions and the *Journal of Applied Mechanics;* to members, $1.00 a copy, $6.00 annually. Add $1.50 for postage to all countries outside the United States, Canada, and Pan American Union. Changes of address must be received at Society headquarters seven weeks before they are to be effective on the mailing list. Please send old as well as new address. . . . By-Law: The Society shall not be responsible for statements or opinions advanced in papers or . . . printed in its publications (B13, Par. 4). . . . Entered as second-class matter March 2, 1928, at the Post Office at Easton, Pa., under the Act of August 24, 1912. . . .

Nonlinear Phenomena

By C. A. LUDEKE,[1] CINCINNATI, OHIO

It is the purpose of this paper to present briefly some phenomena associated with nonlinear systems. The phenomena discussed are wave form, frequency dependence on amplitude, the jump phenomenon, subharmonic oscillations, limit cycles, and frequency entrainment. Experimental examples of such phenomena are included and a representative bibliography is given.

Introduction

FOR the past 15 years it has been the author's pleasure to work in the field of nonlinear mechanics. Although so named (1)[2] this field treats any phenomena, mechanical, electrical, hydrodynamical, and so on, which can be described by nonlinear differential equations. Since most physical phenomena, when accurately represented, require nonlinear differential equations, the task at first glance seems to be to find solutions which are in better agreement with reality and which avoid the idealization of having been linearized. A simple step into the field encourages this belief. For example, consider the equations

$$m\ddot{x} + c\dot{x} + kx = F\cos\omega t \quad [1]$$

$$m\ddot{x} + c\dot{x} + \alpha x + \beta x^3 = F\cos\omega t \quad [2]$$

The first equation is linear and we know the solution to be sinusoidal; the second equation is quasilinear if $\beta x^3 \ll \alpha x$, represents many systems more accurately than Equation [1], and has an almost sinusoidal solution as shown in Fig. 1 (2). Perhaps our task

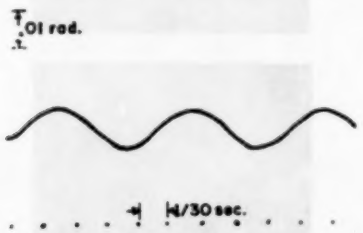


Fig. 1 Analog Solution of Equation [2] With $\beta x^3 << \alpha x$

is "merely" to get better and better solutions by means of higher and higher approximations. This never seemed very exciting to the author and fortunately this is not the case. His suspicions were aroused by Equation [3]

$$m\ddot{x} + c\dot{x} + \beta x^3 = F\cos\omega t \quad [3]$$

which is the same as Equation [2] with $\alpha = 0$. Two things now confound the issue; if you were hoping to use quasilinear methods you now have no linear equation with which to begin, and if you use an analog and expect a radical change between the solutions of Equations [1], [2], and [3], because the mathematics dictates a radical change in procedure, you are again in difficulty as shown in Fig. 2 (3). So although all of the physics must be in the solution, a casual look at them does not tell the story. Perhaps, therefore, we should begin to examine phenomena which show marked contrasts when occurring in linear and nonlinear systems.

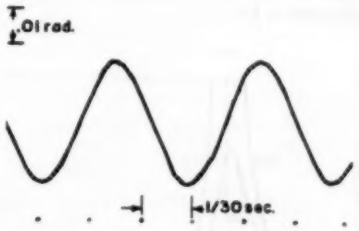


Fig. 2 Analog Solution of Equation [3]

Frequency-Amplitude Dependence

The simplest and most familiar of these is the frequency-amplitude dependence which exists in oscillations occurring in nonlinear systems. Consider systems represented by

$$m\ddot{x} + c\dot{x} + kx = 0 \quad [4]$$

$$m\ddot{x} + c\dot{x} + \alpha x + \beta x^3 = 0 \qquad \alpha > 0 \quad [5]$$

Both of these equations represent damped oscillations. Equation [4] is linear, and as the amplitude of the oscillation decreases the frequency remains unchanged. Equation [5] is nonlinear, and Fig. 3 (4), which gives an analog solution for $\beta > 0$, shows that the frequency decreases as the amplitude decreases. Fig. 4 shows the frequency-amplitude relationships for β equal, less than, or greater than zero. This is the simplest nonlinear characteristic to detect

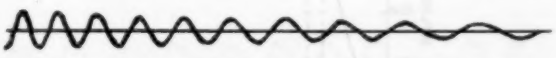


Fig. 3 Analog Solution of Equation [5] With $\beta > 0$
(Note dependence of period on amplitude.)

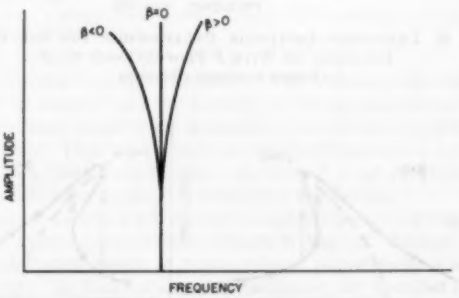


Fig. 4 Frequency-Amplitude Relationship of Various Solutions of Equation [5]
(Free vibrations, linear and nonlinear.)

[1] Department of Physics, Graduate School of Arts and Sciences, University of Cincinnati.

[2] Numbers in parentheses refer to the Bibliography at the end of the paper.

Contributed by the Instruments and Regulators Division of The American Society of Mechanical Engineers and presented at the ASME-AIEE Conference on Nonlinear Control Systems, Princeton, N. J., March 26–28, 1956.

Note: Statements and opinions advanced in papers are to be understood as individual expressions of their authors and not those of the Society. Manuscript received at ASME Headquarters, January 10, 1956. Paper No. 56—IRD-7.

and a graph of the form shown in Fig. 4 tells you not only that nonlinearity is present but also gives an indication of its magnitude.

Jump Phenomenon

If we examine the frequency-amplitude relationship in nonlinear forced oscillations we discover a very striking example of nonlinear phenomena. This is the so-called "jump" phenomenon. Consider once again Equation [1]; it is linear, and the well-known resonance results are shown in Fig. 5. A similar set of results for Equation [2], which is nonlinear, is shown in Fig. 6 (5). Immediately we note that the "resonance curves" are discontinuous and follow different paths for increasing and decreasing frequencies. This is emphasized in Fig. 7.

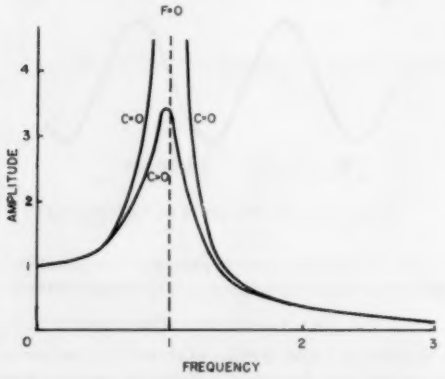


Fig. 5 Frequency-Amplitude Relationship for Solution of Equation [1]
(Forced linear vibrations.)

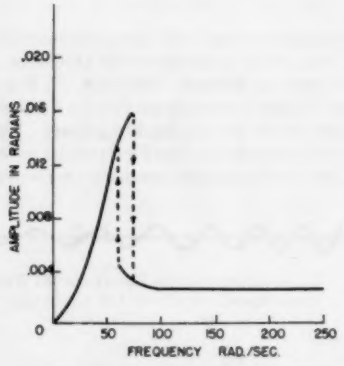


Fig. 6 Frequency-Amplitude Relationship for Solution of Equation [2] With F Proportional to ω^2
(Forced nonlinear vibrations.)

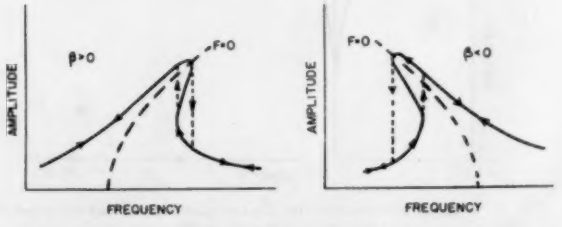


Fig. 7 Jump Phenomena in Frequency-Amplitude Relationships of Forced Nonlinear Vibrations

Not only is Fig. 6 completely different from Fig. 5, but it is also not the only one possible. In considering Equation [1] there is no question about the fact that the solution must have the same frequency as the frequency of the forcing function. However, there is no reason to believe that this also must be true for Equation [2] (6, 7). Actually, solutions exist whose frequencies are multiples or submultiples of the forcing frequency (8). Thus Equation [2] has superharmonic and subharmonic solutions as well as the standard harmonic response. For subharmonics this means that an equation of the form

$$m\ddot{x} + f(x) = F \cos \omega t \ldots\ldots\ldots\ldots\ldots [6]$$

will have as one of its solutions

$$x = A_{1/n} - \cos(\omega t/n) + \ldots + A_1 \cos \omega t + \ldots\ldots [7]$$

Thus the frequency of the fundamental will be a submultiple of the frequency of the forcing function. Specific results from an analog are shown in Fig. 8 (9). A device for demonstrating subharmonic oscillations qualitatively is shown in Fig. 9 (10). We are now in a field of investigation which, with the exception of subharmonics of order one half, is entirely a field of nonlinear phenomena. We can find [for example, see Fig. 10 (11)] that there is a frequency-amplitude relationship for subharmonic solutions. From an actual analog record displayed in Fig. 11 (12)

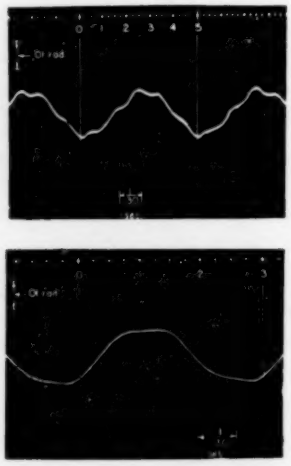


Fig. 8 Subharmonic Wave Forms
(Upper record shows a subharmonic of order one fifth; lower, a subharmonic of order one third.)

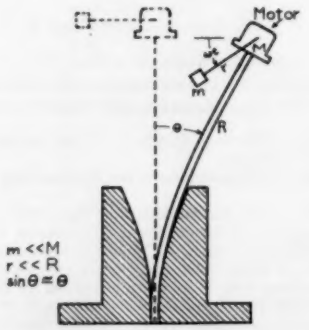


Fig. 9 Device for Demonstrating Subharmonic Vibrations

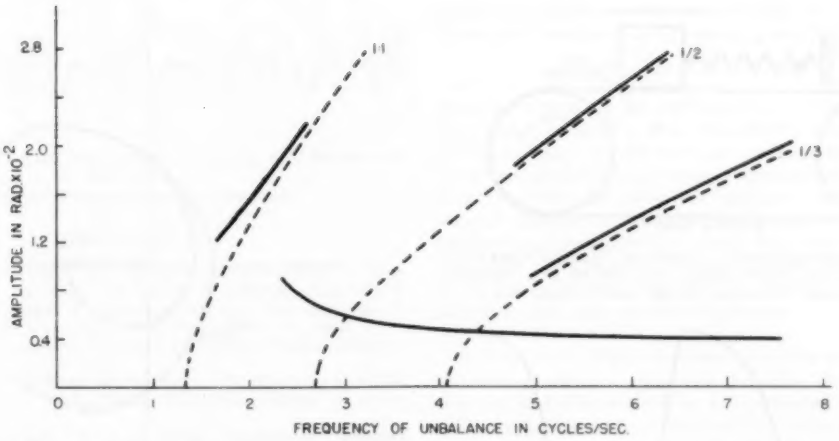


FIG. 10 FREQUENCY-AMPLITUDE RELATIONSHIPS FOR SUBHARMONIC OSCILLATIONS MAINTAINED IN A NONLINEAR SYSTEM BY A ROTATING UNBALANCE
(Vertical jumps take place between the heavy solid lines.)

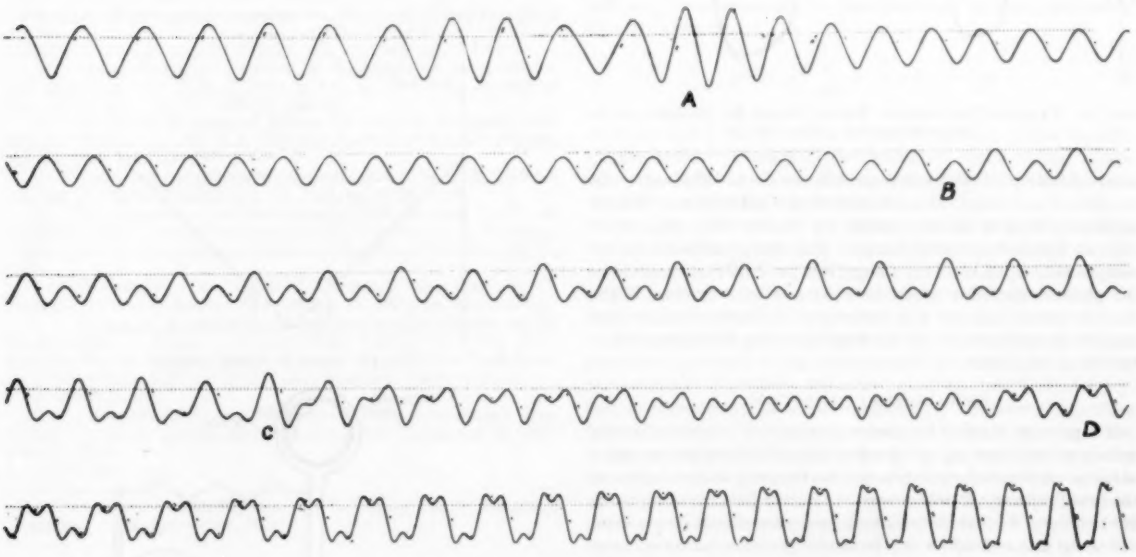


FIG. 11 SUBHARMONIC TRANSITIONS
(At A the jump for harmonic oscillation takes place, and as the frequency increases we have a subharmonic of order 1/2 at B. This subharmonic has its own jump at C, and settles into a 1/3 subharmonic at D.)

we note that transitions may occur between subharmonics of different orders, and that these transitions are associated with a subharmonic jump phenomenon. An entire field is now opened and it is obvious many questions are yet to be answered. Since this is the author's own particular field of interest, he must be on guard against talking too much about it. Let's move on to another nonlinear phenomenon, that of self-excited oscillations.

SELF-EXCITED OSCILLATIONS

Consider an equation of the form (13)

$$m\ddot{x} + \mu(x^2 - 1)\dot{x} + kx = 0 \quad \ldots\ldots\ldots\ldots [8]$$

This equation is nonlinear in the damping term and has no term corresponding to a forcing function. An examination of the nonlinear damping term indicates that for small values of x the damping will be negative and will actually put energy into the system while for large values of x it is positive and removes energy from the system. Thus we are not too surprised that such a system may have an oscillatory solution; and since it is not a forced system we call these oscillations self-excited oscillations.

A system capable of self-excited oscillations is shown in Fig. 12, and the resulting solutions are shown in Fig. 13. Perhaps the most interesting aspect of this nonlinear phenomenon is associated with the final amplitude of oscillation, the so-called limit cycle. This limit cycle depends only on the parameters of the system, not on the initial conditions. This is quite contrary to linear theory. Usually the limit cycles are plotted as closed trajectories in the phase plane as shown in Fig. 14. A common mecha-

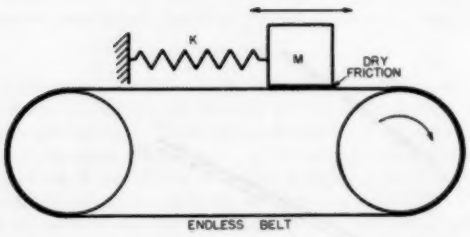


FIG. 12 A MECHANICAL SYSTEM CAPABLE OF CAUSING MASS M TO EXHIBIT SELF-EXCITED OSCILLATIONS BECAUSE OF NONLINEARITY OF DRY FRICTION BETWEEN MASS AND MOVING BELT

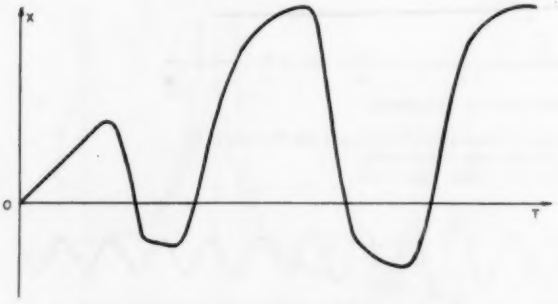


FIG. 13 TYPE OF OSCILLATION WHICH MIGHT BE GENERATED BY SYSTEM SHOWN IN FIG. 12

nism exhibiting a limit cycle is an ordinary clock. This cannot be considered an example of a self-excited oscillation because it is an oscillatory-damped system excited by shocks twice per period from an external source of energy. This energy, released by the escapement, replenishes the energy lost per half cycle and closes the phase trajectories which then become limit cycles. If the clock is wound from rest it is immaterial whether a small or large impulse is used to start it; the final operation is independent of the initial conditions.

ENTRAINMENT OF FREQUENCY

If a periodic force of frequency ω is applied to an oscillating system of frequency ω_0, one observes the well-known phenomenon of beats. As the difference between the two frequencies decreases, the beat frequency also decreases. In a truly linear system, using linear theory, we find that the beat frequency should decrease indefinitely as $|\omega - \omega_0| \to 0$. In actual practice, however, most systems are not truly linear and we find experimentally that the oscillator frequency ω_0 falls in synchronism with, or is entrained by, the external frequency ω within a certain band of frequencies. This phenomenon is called entrainment of frequency and the band of frequency in which entrainment occurs is called the zone of entrainment. The zone of entrainment is indicated in Fig. 15 by the region $\Delta\omega$, which is the region where ω and ω_0 coalesce and only one frequency ω exists. If the theory were linear, the relationship between $|\omega - \omega_0|$ and ω would follow the dotted lines and $|\omega - \omega_0|$ would be zero for only one value of $\omega = \omega_0$. In Fig. 16 we see a system capable of demonstrating frequency entrainment, and in Fig. 17 we see the data obtained from this system in operation.

CONCLUSION

It has been a privilege to discuss some of the phenomena associated with nonlinear systems. There is not too much known about any of these phenomena and each could serve as an excellent field of research, both pure and applied.

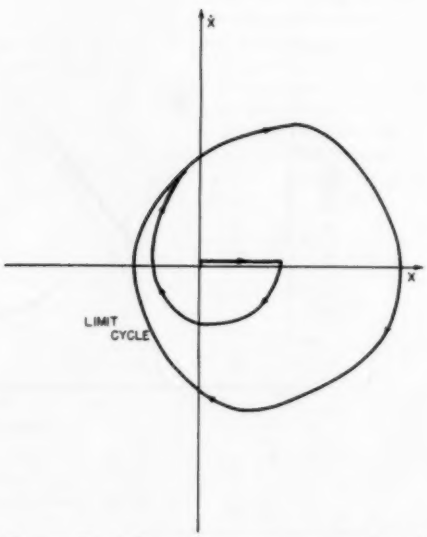


FIG. 14 TYPICAL PHASE-PLANE REPRESENTATION FOR SELF-EXCITED OSCILLATIONS

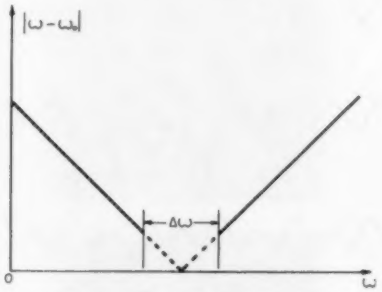


FIG. 15 ZONE OF ENTRAINMENT

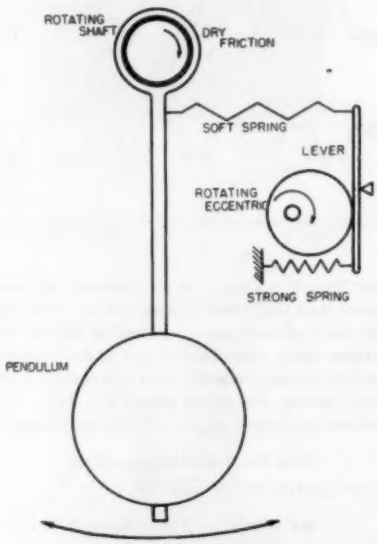


FIG. 16 FROUDE PENDULUM DRIVEN BY A PERIODIC FORCE (Under proper conditions such a system may be used to demonstrate frequency entrainment.)

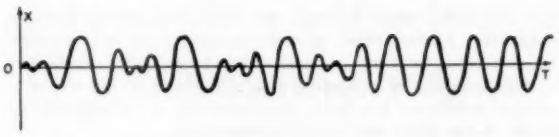


FIG. 17 TYPICAL TIME-DISPLACEMENT CURVE FOR MECHANISM SHOWN IN FIG. 16
(Cessation of beats indicates entrainment.)

BIBLIOGRAPHY

1 Perhaps named after the work entitled "Introduction to Nonlinear Mechanics," by N. Kryloff and N. Bogoliuboff, Kief, USSR, 1937. A free translation of extracts of this work was given by S. Lefschetz, Princeton University Press, Princeton, N. J., 1943.

"Theory of Oscillations," by A. Andronow and S. Chaikin, Moscow, USSR, 1937.

"Introduction to Nonlinear Mechanics," by N. Minorsky and J. W. Edwards, Ann Arbor, Mich., 1947. This book contains an extensive list of references.

"Nonlinear Vibrations," by J. J. Stoker, Interscience Publishers Inc., New York, N. Y., 1950.

"Ordinary Nonlinear Differential Equations," by N. W. McLachlan, Oxford University Press, London, England, 1950.

"Forced Oscillations in Nonlinear Systems," by C. Hayashi, Nippon Printing and Publishing Company, Osaka, Japan, 1953.

2 "An Experimental Investigation of Forced Vibrations in a Mechanical System Having a Nonlinear Restoring Force," by C. A. Ludeke, *Journal of Applied Physics*, vol. 17, 1946, pp. 603–609.

3 "Analog Computer Elements for Solving Nonlinear Differential Equations," by C. A. Ludeke and C. L. Morrison, *Journal of Applied Physics*, vol. 24, 1953, pp. 243–248.

4 "An Electro-Mechanical Device for Solving Nonlinear Differential Equations," by C. A. Ludeke, *Journal of Applied Physics*, vol. 20, 1949, pp. 600–607.

5 "Predominantly Subharmonic Oscillations," by C. A. Ludeke, *Journal of Applied Physics*, vol. 22, 1951, pp. 1321–1326.

6 "The Generation and Extinction of Subharmonics," by C. A. Ludeke, reprinted from Proceedings of the Symposium on Nonlinear Circuit Analysis, Polytechnic Institute of Brooklyn, Brooklyn, N. Y., 1953.

7 "Subharmonic Oscillations in Nonlinear Systems," by C. Hayashi, *Journal of Applied Physics*, vol. 24, 1953, pp. 521–529.

8 "Harmonic, Superharmonic, and Subharmonic Response for Single Degree of Freedom Systems of the Duffing Type," by J. C. Burgess, Technical Report No. 27, Contract N6-ONR-251 Task Order 2 (NR-041-943), Stanford University, Stanford, Calif., 1954.

9 Same as (5).

10 "A Mechanical Model for Demonstrating Subharmonic Resonance," by C. A. Ludeke, *American Journal of Physics*, vol. 16, 1948, pp. 430–434.

11 Same as (5).

12 Same as (5).

13 "On Relaxation Oscillations," by B. Van der Pol, *Philosophical Magazine*, vol. 2, November, 1926.

Discussion

CHIHIRO HAYASHI.[8] Subharmonic oscillations of order 1/2 may occur in nonlinear systems and also in linear systems when system parameters change periodically with time. The device shown in Fig. 9 of the paper has a nonlinear characteristic provided by the guiding support of the pendulum. It also has a time-varying characteristic given by the rotating mass m around the motor. It is rather difficult to see which one of these characteristics (nonlinear or time-varying) gives rise to the subharmonic response. If the guiding surface of the support is removed and still subharmonic oscillations of order 1/2 could be obtained, one could conclude that the subharmonic response is due to the time-varying characteristic of the pendulum.

Generally speaking, subharmonic oscillations of order 1/2 in nonlinear systems are apt to occur when $f(x)$ in Equation [6] is a nonodd function of x, for example, when $f(x)$ is given by

$$f(x) = c_1x + c_2x^2 + c_3x^3$$

However, even when the nonlinearity is symmetrical, i.e., when $f(x)$ is an odd function of x, subharmonic oscillations of order 1/2 may occur. In this case the nonlinearity is symmetrical, but the oscillation itself becomes unsymmetrical and self-biased. In other words, the oscillation contains a unidirectional component. Further, subharmonic oscillations of order 1/2 with different amplitudes and wave forms may occur in the same system under the action of an equal periodic force. This is evidently due to the nonlinearity of the system and cannot be explained by the time-varying characteristic.

K. KLOTTER.[4] The author has given a survey of a number of nonlinear phenomena and, with his often displayed skill, has demonstrated some of them by well-performed experiments.

Among the experiments shown at the Conference (not recorded, however, in the preprints of the paper) was the pendulum with oscillating support. As the author has mentioned, it is well known[5] that such a pendulum may oscillate at a frequency which is half the frequency of the support, and it is equally well known that the phenomenon can be described by a Mathieu differential equation of form

$$\varphi'' + \varphi\left[\frac{g}{l\Omega^2} + \frac{U}{l}\cos x\right] = 0 \quad \ldots\ldots\ldots\ldots [9]$$

where U and Ω denote amplitude and frequency of the support, l is the equivalent length of the pendulum, and $x = \Omega t$. Equation [9] is a linear differential equation with nonconstant (sinusoidally varying) coefficients.

The author then went on to demonstrate very clearly subharmonics of order 1/2, 1/3, 1/5, 1/9, etc., by means of the device shown in Fig. 9 of the paper. The subharmonics in this experiment were all attributed to the nonlinearity in the restoring force of that device.

The writer has some doubts as to the correctness of this explanation in regard to the subharmonics of order 1/2 (or of any order $1/2n$). Certainly, the experiment, as described and performed leaves room for other explanations: Because the exciting force is provided by the inertia force of a single rotating mass, it has not only a horizontal component (as used for explaining the results of the experiment) but also a vertical one. This sinusoidally varying vertical component of the inertia force produces a periodically varying effective stiffness of the spring;[5] an equation of motion of the Mathieu type will follow and hence the situation is quite analogous to the one of the pendulum with oscillating support. The subharmonics observed may be solutions of a (linear) Mathieu differential equation, not solutions of a nonlinear differential equation.

The writer attributes quite a bit of significance to this situation, because in so far as he is aware, the existence of subharmonics of order 1/2 (or of order $1/2n$ generally) in systems having nonlinear restoring forces which are point-symmetrical (represented by purely odd functions of the displacement) has not yet been proved or disproved beyond doubt. Hence it would seem highly desirable to refine the experiments shown by the author in two different ways:

1 The experiment of Fig. 9 should be carried out with a linear spring (omitting the constraining walls and using very small dis-

[8] Department of Mechanical Engineering, Massachusetts Institute of Technology, Cambridge, Mass.

[4] Professor of Engineering Mechanics, Stanford University, Stanford, Calif. Mem. ASME.

[5] "Stabilisierung und Labilisierung durch Schwingungen," by K. Klotter, *Forschung*, vol. 12, September–October, 1941, p. 209.

placements). If the subharmonics of order 1/2 still readily appear (as this writer surmises) the explanation by means of the linear Mathieu equation is the only one applicable.

2 In the nonlinear device of Fig. 9 an excitation should be produced by means of a purely horizontal force (sinusoidally varying). This could be accomplished by the use of two masses rotating clockwise and counterclockwise, respectively, thus destroying their vertical components. Depending on whether or not a subharmonic of order 1/2 will appear in this case one would have an indication about the existence of subharmonics of order 1/2 in systems having odd restoring forces.

The writer would like to urge the author strongly to devote his experimental skills to finding clear-cut answers to this still rather beclouded problem of subharmonics of order 1/2.

Author's Closure

Drs. Hayashi and Klotter both indicate in their discussions that the subharmonic of order one half may arise in systems which have nonlinearities, in systems which have time-varying characteristics, and in systems which have both. Since the system consisting of a rotating unbalance mounted on top of a vertical cantilever has both characteristics, it is difficult to decide which gives rise to the subharmonic.

Both writers suggest an experiment in which the nonlinearity is removed, but not the time-varying characteristic; and Dr. Klotter suggests also the possibility of removing the time-varying characteristics while retaining the nonlinearity. The author is in full agreement with these suggestions, and as soon as time permits he will perform these experiments and any others which suggest themselves in order to locate uniquely the characteristic which gives rise to subharmonic oscillations of order one half.

The author wishes to thank the discussers for many kind words of encouragement and many fine discussions while at the Princeton meeting.

A Résumé of the Development and Literature of Nonlinear Control-System Theory

By T. J. HIGGINS,[1] MADISON, WIS.

"Es gibt nichts Praktischeres als eine gute Theorie (There is nothing more practical than a good theory)" —*Boltzmann*.

The first section of this paper stresses the fact that nonlinear control-system theory is rooted in the theory developed earlier for the solution of other types of nonlinear systems, especially those of nonlinear mechanics and nonlinear electric circuits; whence the control engineer interested in nonlinear systems should gain a thorough knowledge of nonlinear system theory in the large. To facilitate such study this paper advances a concise résumé of the major stages of development of nonlinear system theory to date; a list of papers which encompasses a more detailed account of development; a list of the principal books on general theory; and a list of sources on special aspects of the general theory—each item being characterized as to its particular merit. The principal methods of nonlinear analysis and the phenomena of particular interest in nonlinear systems are noted. Finally, the possible use of these methods and the actual occurrence of these phenomena in nonlinear control-system analysis are emphasized.

I The Roots of Nonlinear Control Theory

THE theory and methods pertinent to the analysis of present-day automatic control systems are essentially the same as the theory and methods developed for the analysis of nonlinear problems encountered in certain domains of nonlinear mechanics and nonlinear electric circuits. In such connection, much of this theory and many of the methods have long been studied, highly developed, and well used for the solution of numerous different kinds of problems. Direct study of most of the some two hundred papers published to date on nonlinear control theory reveals:

1 That essentially all of the theory and methods used to effect analysis of the operating performance of the systems considered in these papers are paralleled in papers published earlier in the other mentioned domains. For example, the basic tenets of the much-employed method of "describing functions" are essentially the same as for Kryloff's method of first approximation, used by him some 20 years ago for the study and solution of various nonlinear electrical and mechanical problems.

2 That a great deal of theory and numerous powerful methods developed for the solution of problems in nonlinear mechanics and nonlinear electric circuits have not yet been employed for nonlinear control-systems analysis—although such use could prove most fruitful.

In such state, it would seem that the control engineer who is interested in obtaining a thorough grounding in theory and methods, which will afford him knowledge both of the causes of various puzzling (to many workers) phenomena occurring in nonlinear control systems and of means of analysis of such systems, could hardly do better than to study and master the theory and methods that have been developed to date in the remarked fields of nonlinear mechanics and nonlinear electric circuits. Such would not prove especially difficult; for, as evidenced in the following:

1 The essential literature is quite circumscribed.

2 Much of the essential content is readily available in English.

3 In most of this content, emphasis is on basic theory, with application illustrated by rather simple physical systems; whence no considerable knowledge of a specific branch of technology is required for ready assimilation.

II The Time Stream of Nonlinear Theory

A desirable first step in undertaking such a study would be to gain perspective of the domain of nonlinear (dynamic) theory as a whole, both in time and in content.

The time stream is clearly marked and consideration thereof tends to a natural division into several stages of progress which may be characterized broadly as follows: The first stage encompasses the initial work, in the eighteenth and much of the nineteenth centuries, on various nonlinear problems whereof solutions could be effected in a specific sense, either in closed form (as in determination of the large oscillations of a pendulum in terms of elliptic functions) or—more usually—in an approximate form, by linearizing the nonlinear equations and obtaining therewith approximations valid over limited ranges of the independent variables (as in the so-called small oscillations of a pendulum).

A second stage comprises the period of, roughly, 1880–1920. On one side it encompasses such powerful and generally applicable developments as Poincaré's topological research (1881–1882), leading to the concept and theory of limit cycles; Linstedt's successful solution (1883) of the secularization difficulty which had baffled Poisson and others in their efforts to effect successful series expansions in periodic terms; Poincaré's work (1892) in effecting analytic solution of a system by means of series expansions in terms of powers of certain numerically small parameters characterizing the nonlinearity; Liapounoff's work (1892) on stability theory; and Bendixsen's investigations (1901) on limit cycles. On the other side, this second stage embraces development of a number of rather specialized procedures enabling approximate solution of special problems—as typified, say, in Martienssen's paper (1910) and in Duffing's book (1918). In the large, this work was concerned with study of problems in nonlinear mechanics and effort, in general, was largely concentrated in this area.

In the third stage, however—from about 1920 to 1940—the most rapid advance in the theory and methods of nonlinear analysis, and in insight into, and understanding of, nonlinear phenomena, stemmed from focus of attention on various phases of nonlinear electric-circuit analysis. Thus this period opened with van der Pol's analysis (1920) of the triode oscillator by his specially developed method. From this work originated a great volume of subsequent associated investigation on electron-tube

[1] Department of Electrical Engineering, University of Wisconsin.

Contributed by the Instruments and Regulators Division of The American Society of Mechanical Engineers and presented at the ASME-AIEE Conference on Nonlinear Control Systems, Princeton, N. J., March 26–28, 1956.

Note: Statements and opinions advanced in papers are to be understood as individual expressions of their authors and not those of the Society. Manuscript received at ASME Headquarters, January 12, 1956. Paper No. 56—IRD-4.

devices. This includes (to cite some of the major workers whose names are yet familiar) that of Appleton and Greaves in England; Liénard, Le Corbeiller, and Rocard in France; van der Pol and van der Mark in The Netherlands; and in the later 1930's and early 1940's, under the stimulus of military communications research, Cartwright and Littlewood in England and Levinson in the United States.

In the later 1920's and through the 1930's, however, the greater volume of work was carried out in Russia. In the large, two groups of workers are to be distinguished. One comprised Mandelstam, Papalexi, Andronow, Chaikin, Witt, and others, centered in Moscow and Leningrad, whose work is largely rooted in Poincaré's analytic and topologic procedures and is supported and complemented by a considerable amount of experimental work. The other comprised Kryloff, Bogoliuboff, and their pupils, at Kiev, whose major work lies in their development of certain linearization procedures and in utilization of Linstedt's analysis, as evidenced by use of these to solve numerous problems in nonlinear mechanics and nonlinear electric circuits.

Although during this stage major attention focused on electron-tube devices, work also was done on nonlinear problems concerned with rotating electric machines (as evidenced in the work of Lyon and Edgerton on pull-out torques of synchronous machines) and on static circuits containing a nonlinear element such as an arc or ferromagnetic inductor (as indicated in Pedersen's work on subharmonic generation in iron-cored circuits).

However, though attention largely centered on problems of nonlinear electric circuits, some work was done on mechanical systems, for example, in vibrating systems encompassing nonlinear springs. Also in this period appears the first work—in a modern sense—on nonlinear control systems as exemplified in Minorsky's early work (1922) on ship steering and in Hazen's later considerations (1934) of relay-type servosystems. In considerable measure publication of Hazen's paper may be said to have initiated the current era of automatic control-system theory. In the large, however, this era really got under way about 1940, under the press of military necessity.

Thus, in the fourth stage of the development of nonlinear system theory, roughly, 1940–1950, a considerable attention centered on nonlinear aspects of control-system analysis. Mainly, the earlier efforts were bent in two directions: To the analysis of systems which are nonlinear by virtue of the deliberate inclusion of a major nonlinear element, as exemplified by the switching action in relay (off-on) servosystems, and to the analysis of systems which are essentially linear, but have undesired nonlinear aspects such as dead time, backlash, elastic hysteresis, saturation, and so on, of a "small" nature. This stage encompasses the work of (to name only some of the best-known non-Russian workers) MacColl, Oldenbourg and Sartorius, Oppelt, Tustin, Kochenburger, Dutilh, Loeb, and Cahen. In Russia, a very considerable volume of work was accomplished by Malkin and many others.

The fifth and present stage of work originated about 1950: On one side, through the start of a considerable effort directed toward improving the performance of control systems by the deliberate inclusion of nonlinear elements and effects (as in MacDonald's work on multiple-mode switching); on the other side, through attempt to study inherent nonlinear effects in control systems in the "large," rather than just in the small wherefor linearization yields workable solutions.

III Orientation

Now, as detailed in Section IV, the major nonlinear phenomena and methods of analysis observed or employed to date in the fourth and fifth stages of the time stream are paralleled, or rooted in, those observed or developed in the first three stages; whence the importance to the control engineer interested in nonlinear systems of gaining a good knowledge pertinent to the first three stages. A good orientation in the broad outlines of this work, through delineation of the major phenomena occurring in nonlinear systems and by means of illustrative physical example without for the most part getting into any considerable mathematics, can be obtained by reading a certain group of papers. These are limited in number, easily obtained, highly readable, and provide a good background for embarking on a detailed study of the theory.

The principal papers of this group, and their particular values, may be mentioned as follows:

1 A comprehensive paper by van der Pol (1)[1] (1934), which gives an excellent chronologically based résumé of the main course of development of nonlinear electric-circuit analysis to that time, primarily as it concerns oscillation problems in radio engineering.

2 A long paper by Mandelstam (2) and others (1935), which outlines the general course of development of nonlinear theory to that time, as developed in Russia by the Moscow-Leningrad group, and as it concerned oscillation problems in engineering.

3 A lengthy paper by von Karman (3) (1940) which, though concerned largely with the development of nonlinear static mechanics, yet contains some account of nonlinear oscillatory problems.

4 Two well-detailed papers by Cartwright (4, 5) (1949, 1952). The first contains a good outline of the course of development of the basic analytic theory and the discovery of the major nonlinear phenomena. The second contains an interesting résumé of the paralleling advances in the United States, England, and Russia in the late 1930's and early 1940's.

5 An interesting paper by Bothwell (6) (1952) on the "present" state of nonlinear stability theory.

6 A paper by Kestin and Zaremba (7) (1952) which provides a good introductory survey of topological analysis and nicely complements item 5.

7 A paper by Weber (8) (1953) of interest in its account of certain nonlinear problems of electric-circuit analysis, aside from those of radio engineering.

8 Minorsky's recent account (9) (1954) of Poincaré's influence on the development of nonlinear theory, as evidenced in the work of certain of the principal investigators to date.

9 Finally, note may be made of certain shorter papers by van der Pol (10) (1948), Cartwright (11, 12) (1948, 1951), and Haacke (13) (1953) which complement the foregoing major items, and of the survey encompassed in the introduction to McLachlan's book, mentioned later.

IV Detailed Accounts of General Theory

With background thus gained on the course of development to date, the engineer interested in obtaining a firm grounding in nonlinear control theory should next turn to study of several (and subsequently all—if time permits) of the half-dozen principal collective treatments on nonlinear analysis published to date in English. As a guide to such study, the following brief characterization may be of interest:

1 N. Kryloff and N. Bogoliuboff (14) (1943). This is a free translation and condensation of material from two monographs by these authors in the 1930's. It provides a good introduction to both their methods based on Linstedt's analysis and to certain procedures of linearization based on consideration of the physical aspects of a system.

2 A. Andronow and S. Chaikin (15) (1949). This is a free

[1] Numbers in parentheses refer to the Bibliography at the end of the paper.

translation and condensation of the 1937 Russian edition. It provides an introduction to topological procedures, to van der Pol's and Poincaré's methods, to solution by piecewise linearization, to graphical solution by isoclines, and to other aspects of the general theory.

3 N. Minorsky (16) (1947). This is essentially a reissue in one volume of four 1944–1945 David Taylor Model Basin Reports. Primarily, this text comprises a connected account of Russian work as evidenced in items 1 and 2. It encompasses much material omitted in these English translations and condensations, but it also omits some contained therein (especially in item 1). The text is supported throughout by discussion of numerous illustrative examples. A second, considerably expanded, edition is now in preparation. A good résumé of much of the work discussed in this book is provided in Minorsky's (17) interesting paper, "Modern Trends in Nonlinear Mechanics" (1948); see also Bellin's (18) paper, "Non-Autonomous Systems" (1953).

4 N. W. McLachlan (19) (1955). This enlarged second edition of the original 1950 text provides a good mathematical introduction to numerous analytic procedures.

5 J. J. Stoker (20) (1950). A good introductory text, shorter and less detailed in analysis than is item 4.

6 C. Hayashi (21) (1953). This text is devoted largely to exposition of forced oscillations (whereas the previous items are largely concerned with free oscillations) and encompasses accounts of corresponding experimental investigation and substantiation of various phases of the theory.

V Special Aspects of the General Theory

Various special aspects of the general body of nonlinear theory are entailed in readily available published books and collections of notes. These may prove of interest and aid as noted:

1 H. Poincaré (22) (1892). Poincaré's own account of his method of small parameters is yet one of the best.

2 A. Liapounoff (23) (1947). The 1947 photoreproduction of the 1907 corrected French translation of the original 1892 work comprises a detailed account of one of the central features of nonlinear-oscillation theory, namely, Liapounoff's stability criteria. In this connection, note Section III—5.

3 C. Duffing (24) (1918). This book comprises a detailed study of what is now known as Duffing's equation.

4 P. Le Corbeiller (25) (1931). This contains a descriptive account of numerous nonlinear phenomena and of physical systems in which they occur.

5 W. Hurewicz (26) (1943). These mimeographed notes provide an excellent introduction to certain topological aspects of the solutions of nonlinear problems

6 Friedrichs, Stoker, Le Corbeiller, and Levinson (27) (1942–1943) and Eckweiler, Flanders, Stoker, Friedrichs, and John (28) (1946). Much of the essential content of these mimeographed notes, of courses given, respectively, at Brown and at New York University, are encompassed in Stoker's (20) book.

7 L. Pipes (29) (1946). In this book, chapter 22, "The Analysis of Nonlinear Oscillatory Systems," contains a good rephrasing of some aspects of Kryloff-Bogoliuboff's theory of higher approximation in terms of Laplace transform theory.

8 A. Madwed (30) (1950). This recently reprinted report provides the best available general account of the number-series method of solution, initiated largely by Tustin.

9 Flügge-Lotz (31) (1953). This book stems from reports written in Germany during the last war in connection with design of control systems for guided missiles. It provides a good connected account of study of forced relay-type servo-systems by phase-space methods.

10 R. Bellman (32) (1953). This text contains an interesting account of certain aspects of the stability of nonlinear systems, as in chapter 4 on Poincaré's and Liapounoff's work. It encompasses much of an earlier research report of essentially the same title.

VI Principal Methods of Nonlinear Analysis

The books of Section IV on general theory and the complementary specialized items of Section V provide, at best, only a basic introduction to some of the better-known and most-used methods of nonlinear analysis. After study of these texts, one who desires to know more of the advanced phases of the theory and uses of these methods, and of a considerable body of other theory and numerous useful other methods, would next turn to the periodical literature. This is very extensive, indeed. If, however, attention is directed to that which is of especial value in regard to nonlinear control-systems work, a considerable limitation results. In fact, most of the consequential work is encompassed in a few hundred titles.

Lack of space forestalls discussion of even this selected list. It must suffice to remark that over the past several years the writer has compiled a rather exhaustive bibliography of the major published work on nonlinear analysis, as applicable for investigation of engineering problems; has read the greater portion of this body of work; and is preparing a classified listing of those papers which cover theory and methods of analysis of interest in nonlinear control-system work.

However, indicative of the very considerable number and variety of such methods, many of which have not yet been employed in nonlinear control-system analysis, it is of value to enumerate here the principal available procedures (analytic, topologic, graphic, numeric, and computer-effected). Thus:

1 Solution of the equations of performance in closed form in terms of known functions, such as elliptic or hyperelliptic functions.

2 Sectional linearization (yielding linear differential equations of performance holding over stated ranges of the independent variables).

3 Equivalent linearization, as effected by energy-balance, etc.

4 Expansion in powers of one or more parameters characterizing small nonlinearities, as in Poincaré's method.

5 Kryloff and Bogoliuboff's methods of first approximation, improved first approximation, and higher approximation as based on Linstedt's procedure.

6 Van der Pol's method (which has much in common with the K-B method of first approximation, the former being couched in variable polar co-ordinates, whereas van der Pol's is couched in variable rectangular co-ordinates).

7 Minorsky's stroboscopic method.

8 Series expansion in powers of the input function.

9 Inversion of series.

10 Fourier expansions (for investigating steady-state oscillations)

11 Asymptotic-series expansions.

12 Various methods of rather limited usefulness, such as, Duffing's method, one-term approximations to a solution, and so on.

13 Use of nonlinear integral equations.

14 Transform theory, enabling a simpler effectance of the analysis by the methods of items 5, 9, and 13.

15 Various variational procedures, especially Galerkin's method and Ritz's method.

16 Various iterative procedures.

17 Number-series method.

18 Various numerical procedures, especially as formulated by Collatz.

19 Various aspects of phase-plane and phase-space analysis (topological procedures).

20 Numerous methods of graphical construction in the phase plane and phase space: Isoclines, Liénard's, arc-segment procedures, and so on.

21 Graphical constructions of rather limited usefulness, such as Rauscher's or Martiennsen's methods.

22 Miscellaneous procedures: Variation of parameters, successive integrations, Picard's method of successive approximations, method of collation, and so on.

23 Analog-computer and digital-computer solution.

24 Direct modeling in miniature.

VII Phenomena of Particular Interest in Nonlinear Systems

Detailed study of the body of work discussed in Sections V and VI evidences that the major considerations in design or analysis of nonlinear systems are largely encompassed under one (or more) of the following:

1 Determination of free or forced transient response of a system with stated parameters and prescribed initial conditions.

2 Determination of free or forced steady-state response of a system with stated parameters and prescribed initial conditions, and, when a response is oscillatory, how the amplitude, frequency, period, and nature of this cyclic response depends on the stated parameters and initial conditions.

3 Determination of bounds on the amplitude of steady-state cyclic response.

4 Determination of the nature of the stability of steady-state cyclic response.

5 Establishment of conditions enabling knowledge of the degree of accuracy to be expected from application of various procedures for effecting approximate determinations of steady-state, free or forced cyclic response, as by Martienssen's or Schwesinger's methods.

6 Determination of subharmonic and multiple-harmonic steady-state cyclic responses, particularly if these dominate the fundamental in amplitude.

7 Determination of interaction effects between two or more elements in a nonlinear system.

8 Determination of jump phenomena, characterized by sudden increase in the value of a system response, expressed as a function of a parameter or an input function.

9 Determination of various resonance phenomena, characterized by unusually large amplitudes of response stemming from certain critical combinations of parameters and/or initial conditions.

10 Investigation of the causes, and means of causing or obviating, autoexcitation, parametric excitation, and other such effects.

VIII Nonlinear Automatic Control Systems

In substantiation of the values of the foregoing program for obtaining a good grounding in the theory and methods of nonlinear analysis, it is to be remarked that a detailed study of the literature published to date on nonlinear control systems reveals:

1 That the various phenomena studied therein are encompassed among those just listed in Section VII.

2 That essentially all of the analysis utilized for solution or study of specific nonlinear control problems is rooted in, or paralleled by, methods stated in Section VI.

This body of published literature on nonlinear control systems is not prohibitively large. It comprises rather sketchy treatment in several general texts in English (the most comprehensive is, perhaps, that in Truxal's (33) book); Flügge-Lotz's (31) text; several Russian books (detailed in the author's (34) recent paper); and a body of several hundred periodical papers. Accordingly, it is not difficult for the engineer especially interested in nonlinear control systems to master what has been published. As an aid to such effort, the author has prepared a classified listing of much of the consequential, non-Russian periodical literature on nonlinear control-system theory published to date. A copy may be obtained upon request.

BIBLIOGRAPHY

1 "The Non-Linear Theory of Electric Oscillations," by B. van der Pol, Proc. Inst. Radio Eng., vol. 22, 1934, pp. 1051–1086.

2 "Exposés des récherches recentes sur les oscillations nonlinéaires," (Account of Recent Researches on Nonlinear Oscillations) by L. Mandelstam, N. Papalexi, A. Andronow, S. Chaiken, and A. Witt, *Tech. Phys.*, USSR, vol. 2, 1935, pp. 81–134.

3 "The Engineer Grapples With Nonlinear Problems," by T. von Karman, *Bull. Am. Math. Society*, vol. 46, 1940, pp. 615–683.

4 "Non-Linear Vibrations," by M. L. Cartwright, Adv. Sci. (British Asso. Adv. Sci.), April, 1949, pp. 1–12.

5 "Non-Linear Vibrations: A Chapter in Mathematical History," by M. L. Cartwright, Math. Gaz., vol. 36, 1952, pp. 81–88.

6 "The Current Status of Dynamic Stability Theory," by F. E. Bothwell, Trans. AIEE, vol. 71, part 1, 1952, pp. 223–228.

7 "Geometrical Methods in the Analysis of Ordinary Differential Equations: Introduction to Nonlinear Mechanics," by J. Kestin and S. K. Zaremba, *Applied Science Research*, series B, vol. 3, 1952, pp. 149–189.

8 "Introduction to Nonlinear Physical Phenomena," by E. Weber, Proc. Sym. Nonlinear Circuit Analysis, New York, N. Y., 1953, pp. 1–27.

9 "Influence d'Henri Poincaré sur l'evolution moderne de la theorie des oscillations nonlinéaires" (Influence of Henri Poincaré on the Modern Evolution of the Theory of Nonlinear Oscillations), by N. Minorsky, Mem. Soc. Ingrs. Civ. Fr., vol. 107, no. 111, July–September, 1954, pp. 108–205.

10 "Mathematics and Radio Problems," by B. van der Pol, Philips Res. Report 3, 1948, pp. 174–190.

11 "Topological Aspects of Forced Oscillations," by M. L. Cartwright, *Research*, vol. 1, 1948, pp. 601–606.

12 "Non-Linear Vibrations," by M. L. Cartwright, *Journal Pi Mu Epsilon*, April, 1951, pp. 131–137.

13 "Über die nichtlineare Mechanik" (On Nonlinear Mechanics), by W. Haacke, *Phys. Blatter*, vol. 9, no. 9, 1953, pp. 398–405.

14 "Introduction to Non-Linear Mechanics," by N. Kryloff and N. Bogoliuboff, tr. by S. Lefschetz, Princeton University Press, Princeton, N. J., 1943, 105 pp.

15 "Theory of Oscillations," by A. Andronow and S. Chaikin, tr. by N. Goldowski, Princeton University Press, Princeton, N. J., 1949, 358 pp.

16 "Introduction to Nonlinear Mechanics," by N. Minorsky, J. W. Edwards, Ann Arbor, Mich., 1947, 464 pp., second edition pending publication.

17 "Modern Trends in Nonlinear Mechanics," by N. Minorsky, *Advances in Applied Mechanics*, Academic Press, Inc., New York, N. Y., vol. 1, 1948, pp. 41–103.

18 "Non-Autonomous Systems," by A. I. Bellin, *Advances in Applied Mechanics*, Academic Press, Inc., New York, N. Y., vol. 3, 1953, pp. 295–320.

19 "Ordinary Non-Linear Differential Equations in Engineering and Physical Sciences," by N. W. McLachlan, Clarendon Press, Oxford, England, first edition, 1950, 201 pp.; enlarged second edition, 1955.

20 "Nonlinear Vibrations in Mechanical and Electrical Systems," by J. J. Stoker, Interscience Publishers, Inc., New York, N. Y., 1950, 273 pp.

21 "Forced Oscillations in Non-Linear Systems," by C. Hayashi, Nippon Printing and Publishing Company, Osaka, Japan, 1953, 164 pp.

22 "Les méthodes nouvelles de la mécanique céleste" (The New Methods of Celestial Mechanics), by H. Poincaré, Paris, France, 1892.

23 "Problème general de la stabilité du mouvement" (The General Problem of the Stability of Motion), by A. Liapounoff, Annals of Mathematical Studies, no. 17, Princeton University Press, Princeton, N. J., 1947.

24 "Erzwungene Schwingungen bei verandlicher Eigenfrequenz"

(Forced Oscillations With Arbitrary Eigentrequency), by C. Duffing, Braunschweig, Germany, 1918.

25 "Les systèmes autoentretenus et les oscillations de relaxation" (Self-Oscillating Systems and Relaxation Oscillations), by P. Le Corbeiller, Librairie Scientifique Hermann et Cie, Paris, France, 1931, 46 pp.

26 "Ordinary Differential Equations in the Real Domain With Emphasis on Geometric Methods," by W. Hurewicz, Brown University, Providence, R. I., 1943, 129 pp.

27 "Nonlinear Mechanics," by K. Friedrichs, J. Stoker, P. Le Corbeiller, and N. Levinson, Brown University Press, Providence, R. I., 1942–1943.

28 "Studies in Nonlinear Vibration Theory," by H. Eckweiler, D. A. Flanders, J. Stoker, K. Friedrichs, and F. John, Inst. Math. Mech., New York University, New York, N. Y., 1946.

29 "Applied Mathematics for Engineers and Physicists," by L. Pipes, McGraw-Hill Book Company, Inc., New York, N. Y., 1946—see chapter 22 "The Analysis of Nonlinear Oscillatory Systems."

30 "Number Series Method of Solving Linear and Nonlinear Differential Equations," by A. Madwed, Report No. 6445-T-26, Instrumentation Laboratory, Massachusetts Institute of Technology, Boston, Mass., 1950.

31 "Discontinuous Automatic Control," by I. Flügge-Lotz, Princeton University Press, Princeton, N. J., 1953, 168 pp.

32 "Stability Theory of Differential Equations," by R. Bellman, McGraw-Hill Book Company, Inc., New York, N. Y., 1953, 166 pp.

33 "Automatic Feedback Control System Synthesis," by J. G. Truxal, McGraw-Hill Book Company, Inc., New York, N. Y., 1955, 675 pp.

34 "Basic Books for Your Control Engineering Library, II. Servomechanisms," by T. J. Higgins, *Control Engineering*, vol. 1, December, 1954, pp. 48–51. A complementary article is to appear in a 1956 issue of *Control Engineering*.

Discussion

RICHARD BELLMAN.[3] It may be worth while, as an appendix to this very valuable summary of the classical and modern theory of nonlinear differential equations and its interrelation with electronic and mechanical circuit theory, to discuss briefly some recent work which has been done by the writer and collaborators in connection with control processes.

We may consider every scientific theory to have several stages of development. In the first stage we attempt to describe and predict physical phenomena arising from given systems; in the second stage we attempt to design systems so as to produce desired phenomena. Although, at first glance, it might seem that a thorough mastery of the first stage is essential to progress in the second stage, this is not necessarily true. In other words, some classes of optimization problems may be easier to tackle than related classes of purely descriptive problems. The reason for this lies in the fact that variational properties possessed by optimal systems in many cases greatly simplify the analytic structure of the equations which describe the system.

A class of mathematical problems which arise in economic as well as engineering context is the following: Consider a nonlinear vector system of differential equations

$$\frac{dx}{dt} = g[x, f(t)], \quad x(0) = c \ldots\ldots\ldots\ldots [1]$$

where $x(t)$ represents the state of a system at time t and $f(t)$ is a control vector which is to be chosen so as to have the system perform in some desired way. An example of some interest in recent years is that of "on-off" or "bang-bang" control, where the components of $f(t)$ assume only values ± 1 and are to be chosen so as to minimize the time required for the system to return to the equilibrium state $x = 0$.

If the system is linear inhomogeneous

$$\frac{dx}{dt} = Ax + f(t), \quad x(0) = c \ldots\ldots\ldots\ldots [2]$$

the problem may be treated in various ways and explicit solutions obtained (cf. Bellman, Glicksberg, Gross,[4] where other references may be found).

If the system is nonlinear, variational problems arise which escape classical techniques in general. In a number of cases, however, they may be resolved computationally, with the aid of modern high-speed computers, using the techniques of the theory of dynamic programming, Bellman,[5,6] and the new approach to the calculus of variations furnished by this theory.

A discussion of some of the applications of this method to the general problem discussed in the foregoing is contained in Bellman.[7]

K. G. BLACK.[8] This is an interesting and informative paper. It should be of considerable value to all engineers who encounter control problems in which nonlinearities are present. Many control engineers are familiar with certain nonlinear techniques which they have found useful in their work, and the excellent list of nonlinear methods presented in the paper should facilitate the selection of others. The author's considerable experience in teaching courses in both nonlinear and control theory has enabled him to suggest a program of study which offers both an aid and a challenge to the control engineer.

T. H. CHIN.[9] To the analytical study and application of the theory of nonlinear control systems, the author makes a most welcome and informative contribution by presenting a concise integrated summary of the progress and of the development of nonlinear control theory to date.

The author emphasizes the possibility of employment of nonlinear theory and techniques to enable design of control systems which yield maximal performance. He provides a most valuable outline of how to go about obtaining the knowledge of this theory.

In recent years, as is well known, the determination of the performance of a control system has been based largely on frequency-response analysis and transient-response analysis. These techniques are most useful for analyzing highly linear systems. However, when a very high degree of accuracy is desired, then, because most physical systems are wholly linear, appropriate attention must be given to the system nonlinearities resulting from backlash, starting friction, hydraulic leakage, thermal expansion, dielectric breakdown, amplifier distortion, resistor thermal dependency, and many other effects.

It has been shown both analytically and experimentally that the utilization of such simple nonlinear theory as phase-plane analysis can embrace certain nonlinear effects in second-order feedback-control systems and lead to substantial improvement in performance. Evidently a very broad knowledge of nonlinear

---

[3] Rand Corporation, Santa Monica, Calif.

[4] "On the Bang-Bang Control Problem," by R. Bellman, I. Glicksberg, and O. Gross, *Quarterly of Applied Mathematics*, vol. 14, 1956, pp. 11–18.

[5] "The Theory of Dynamic Programming," by Richard Bellman, *Bulletin of the American Mathematical Society*, vol. 60, 1954, pp. 503–516.

[6] "Principles of Dynamic Programming," by Richard Bellman, Princeton University Press, Princeton, N. J., in press.

[7] "On the Application of Dynamic Programming to the Study of Circuit Analysis," by Richard Bellman, Symposium on Nonlinear Processes, Brooklyn Polytechnic Institute, Brooklyn, N. Y., April, 1956.

[8] General Electric Company, Schenectady, N. Y.

[9] Assistant Professor, Department of Electrical Engineering, University of Pittsburgh, Pittsburgh, Pa.

theory, such as outlined by the author, would enable corresponding successful study of much more complex systems. Another powerful tool in dealing with nonlinear problems is the use of electrical analogies in which nonlinear effects due to mechanical, thermal, hydraulic, and pneumatic components can be solved conveniently by studying analogous electrical systems through the aid of analog computing devices. It may be noted that detailed information on one form of electrical-analog techniques is given in the author's paper.[10]

Certainly, in dealing with nonlinear problems, adequate possession of the knowledge outlined by the author will often enable a satisfactory solution of nonlinear control problems, which otherwise would be unsolvable.

R. N. Clark.[11] There is little question that nonlinear analysis will continue to be of increasing importance in the control-system industry. There are many engineers now in industry who will have to become schooled in nonlinear mechanics without the benefit of formal education. The definition and identification of the various aspects of this subject that are presented in the paper should prove invaluable in orienting those engineers. The author is to be congratulated on a splendid presentation of a comprehensive job of library research and subject classification.

G. H. Fett.[12] The author has carefully classified the literature of nonlinear control-systems theory. Of special interest is his catalog of methods of analysis.

To his list of 24 methods of analysis that he gives in Section VI several combinations of interest to servomechanism theory can be added. Topological analysis represents a fruitful area of study when supplemented by the use of computers, a combination of Nos. 19 and 23. For example, the operation of the control system, dynamically, can be envisioned as the trajectory in a phase space which gives the error displacement, velocity, and acceleration for a third-order system, and this trajectory can be analyzed. Now, if it is desired to bring the system back to equilibrium, then the distance from the trajectory to the origin should be reduced to zero in the shortest period of time[13] with due consideration to the abilities of the prime mover. With the aid of a linear transformation and a computer[14] a method for obtaining this characteristic can be devised.

Another combination method of attack is represented by the use of power-series expansion of the nonlinear elements in terms of one or more of the variables, a modification of No. 8, coupled with the Fourier expansions of No. 10. This method[15] is particularly useful in establishing average and instant instability criteria in multiple-loop systems.

Equivalent linearization, No. 3, can be used to great advantage to predict transient performance of systems with specific types of nonlinearities such as backlash, saturation, and hysteresis.[16]

A method of synthesis can be extracted from the methods listed under Nos. 23 and 24. A nonlinear control system is set up on a computer with some particular type of nonlinearity. The output is then observed as the input is varied, as the nonlinearity is varied, and as other parameters are changed. Then the solution to the nonlinear system is known, and if the resultant solution is similar to the desired one, a basis for synthesis is developed. A catalog of solutions thus is used as a reference for design.

As an added item to the bibliography which Professor Higgins has provided, the writer would suggest the addition of Volumes II and VI of the MRI Symposia Series published from the papers given at New York meetings of the Professional Group on Circuit Theory of the IRE, in 1953 and 1956.

G. F. Forbes.[17] With reference to item 23, section VI, of the author's paper, commercial digital differential analyzers, (DDA), are now available in the price class of about ten-thousand dollars. Such machines are mathematically equivalent to the mechanical differential analyzer. They have much greater capacity and all of the convenience with respect to the nonlinear terms in the differential equation.

Availability of such machines, on a production basis, indicates a potential radical change in the philosophy and approach to nonlinear system design. In many cases, the cost of the DDA is less than the cost of building a trial model of the nonlinear system. Any nonlinear device for which there is any possibility of an analytic solution can be simulated easily on the DDA. Nonlinear systems, for which no analytic approach is available, usually offer no difficulty to exact simulation.

J. D. Graham.[18] In this paper the author makes a definite contribution to the young worker in the field of nonlinear control systems. By reading this résumé the person not versed in this subject can survey the field with a minimum reading and have an over-all picture of the methods and applications of nonlinear analysis. The paper also makes the suggestion that work is still to be done in applying analysis to control systems that already have been developed for nonlinear mechanics and electrical circuits.

In reading the material presented it should be remembered that the present state of the nonlinear-analysis art is not wholly satisfactory for all the methods presented have serious limitations such as accuracy, length of calculations, and area of application. Thus workers in nonlinear control systems are still challenged to effect new and more satisfactory methods in addition to using methods known in other fields.

J. D. Horgan.[19] To the engineer who has embarked on a program of self-study in nonlinear control-system theory, or who plans to begin such an effort soon, the prospects of finding and evaluating the countless references on the subject is overwhelming, perhaps completely discouraging. Thus it is that this paper, which comprises, in fact, a plan of study, should be of great value. By giving a brief sketch of the historical background, theory, problems, and methods of solution, the author has provided a framework to which future studies may be referred, enabling the engineer to make an integrated approach to the subject. By directing the reader's attention to certain key papers and books, the author has indicated how the limited time available for such studies can be used to best advantage.

The emphasis in this presentation on historical background

[10] "Electronnalogic Methods," by T. J. Higgins, *Applied Mechanics Reviews*, vol. 9, 1956, January, pp. 1–4; February, pp. 49–55.

[11] Research Engineer, Aeronautical Division, Minneapolis-Honeywell Regulator Company, Minneapolis, Minn.

[12] Department of Electrical Engineering, University of Illinois, Urbana, Ill.

[13] "Metrization of Phase Space and Nonlinear Servo Systems," by C. L. Kang and G. H. Fett, *Journal of Applied Physics*, vol. 24, 1953, p. 38.

[14] "Higher Order Servos With Nonlinear Computer Using Phase Space Techniques," by E. J. Hagin, PhD thesis, University of Illinois, 1956, to be published.

[15] "A Phasor Method of Nonlinear Analysis," by J. P. Neal, PhD thesis, University of Illinois, 1955, accepted for publication by the American Institute of Electrical Engineers, paper 56-804.

[16] "Feedback Control Systems," by G. H. Fett, Prentice Hall, Inc., New York, N. Y., 1954, pp. 330–345.

[17] Industrial Mathematician, Pacoima, Calif.

[18] Assistant Professor, Department of Electrical Engineering, Kansas State College, Manhattan, Kans.

[19] Assistant Professor in Electrical Engineering (on leave, Marquette University, Milwaukee, Wis.); at present, Graduate Student, University of Wisconsin, Madison, Wis.

should be noted by all engineers, but particularly by engineering teachers. It is feared that this is a facet of engineering which is too often neglected in the classroom. Yet, it can give life and meaning to a topic which otherwise might appear to be drab and unimportant.

Y. H. Ku.[20] As nonlinear analysis has gone through a great deal of development in the recent years, the author's résumé is most welcome. To quote van der Pol from his discussion on "Some New Concepts and Theorems Concerning Nonlinear Systems," by Cherry and Millar:[21]

"After 1920 technicians soon became interested in these nonlinear differential equations so that soon the whole field was studied intensively in the U. S. A. and there the practical importance of these nonlinear differential equations was soon recognized. Then the war came. I was in Holland all the time during the Occupation so that we did not hear what happened outside in the civilized world, but after the war the door was opened and I was happily surprised to note that especially in France, Great Britain, and also in the United States, a large amount of work had been done and was being done by pure mathematicians during this time in the field of nonlinear differential equations."

Minorsky's paper on "Nonlinear Control Systems"[21] made the following comments on the van der Pol equation:

"The connection between this discovery and the earlier work of Poincaré was ascertained only nine years later. From that moment on the progress in this new field, nonlinear mechanics, became systematic and rapid. Between 1930 and 1939 these studies were conducted almost exclusively in the USSR but after the end of the last war they were taken up also in the United States and in England."

Reference (17) of the paper contained 44 references. Keller's 1941 AIEE paper[22] contained a bibliography of a hundred entries. To these additional references, the writer likes to supplement the contributions edited by Lefschetz[23] and the writer's recent paper[24] on the phase-space method.

R. E. Kuba.[25] The author has provided a valuable and comprehensive paper. It is interesting to note that even though the beginnings of nonlinear control theory date back nearly a century, one may still encompass the entire body of consequential nonlinear control literature by consulting only a few score major works. This should act as a stimulus to bring new workers into a field loaded with research and development possibilities.

Compiling an accurate bibliography and evaluating its basic characteristics is all too often a thankless task; therefore, on behalf of the nonlinear control people, the writer wishes to thank the author for the great deal of personal time and effort he has devoted to this end.

Rolf W. Lanzkron.[26] In this paper the author shows very clearly that the techniques and methods of analysis used for solving problems in nonlinear mechanics, and especially, nonlinear electric-circuit theory, provide powerful means of solving problems in the field of nonlinear control systems, and that there are many such means which have not yet been used in control work.

The author is to be commended for his excellent bibliography. It will be an aid to both the young engineer who is working in the field of nonlinear problems and wishes to get acquainted in general with some of these methods of solution, and the engineer long in the field who is trying to solve a specific problem. This is because the author has given a bibliography in which not only the names of specific books and the locations of certain articles are given, but also a useful discussion of the problems therein, the methods used in their solution, and the special values of each book or article

N. Minorsky.[27] The author is to be congratulated for his presentation of the whole field of nonlinear problems and methods in such a short and, at the same time, clear manner. No less important is his statement regarding the advisability of basic studies in the theory of nonlinear differential equations for those who desire to work on nonlinear-control problems.

Very often one observes, particularly on the part of beginners, a tendency to consider nonlinear problems as something which merely requires a correction in the solutions of the corresponding linear problems. This tendency is due undoubtedly to the existing linear theory of control systems which yields itself to arguments of this kind inasmuch as a continuous variation of co efficients of a differential equation generally accounts for a continuous variation of its solutions. This circumstance found a widespread application in all kinds of diagrams of transfer loci obtained owing to the use of Laplace's transform. In the nonlinear field all this becomes very doubtful and frequently even entirely wrong. The problems are full of pitfalls if looked at from the viewpoint of the familiar linear theory. A certain phenomenon may either appear or disappear abruptly if a certain critical value of a parameter in the differential equation is reached. If one tries to analyze the effect (i.e., the solution of the differential equation), it is impossible to understand what happens and still less to master the phenomenon whose cause is concealed in the properties of the differential equation. The linear tools like Laplace's transform, so useful in the theory based on linear differential equations, cease to be applicable in the nonlinear field. Likewise, the Nyquist diagram of stability does not apply any more, inasmuch as one has to use in the nonlinear field criteria associated with the variational equations of Poincaré, and so on.

All this is due to the fact that the existing theory of control is based (sometime in a rather implicit form) on the theory of differential equations in the complex domain, while the whole nonlinear field belongs to differential equations in the real domain. Or, as is well known, beginning with singularities, everything is different in these two domains and there are no means whatever to establish any connection between them.

The only way out from this situation is to start from the fundamentals, as the author suggests, and build the theory of nonlinear control systems on the nonlinear basis. This may appear somewhat disappointing for those who have acquired a habit of using the existing control theory, but such situations are inevitable in any theory when it has to be overhauled under the impact of new experimental facts.

In this particular case the situation is not so difficult as it may appear from an offhand consideration. In fact, there are many results from the theory of oscillations which can be transferred into the theory of nonlinear control systems. Thus, for instance, an engineer dealing with the theory of oscillations may be interested in establishing conditions for the existence of a limit cycle (i.e., stationary oscillation). A control engineer, on the other hand,

[20] Moore School of Electrical Engineering, University of Pennsylvania, Philadelphia, Pa.

[21] "Automatic and Manual Control," papers contributed to the Conference at Cranfield, 1951, Butterworths Scientific Publications, London, England, 1952.

[22] "Analytical Methods of Solving Discrete Nonlinear Problems in Electrical Engineering," by E. G. Keller, Trans. AIEE, vol. 60, 1941, pp. 1194–1200.

[23] "Contributions to the Theory of Nonlinear Oscillations," edited by S. Lefschetz, Princeton University Press, Princeton, N. J., vols. 1 and 2, 1950 and 1952.

[24] "Analysis of Nonlinear Systems With More Than One Degree of Freedom," by Y. H. Ku, *Journal of The Franklin Institute*, vol. 259, 1955, pp. 115–131.

[25] Associate Professor, Department of Electrical Engineering, Wayne University, Detroit, Mich.

[26] Remington-Rand Univac, St. Paul, Minn.

[27] Professor Emeritus, Stanford University, Stanford, Calif.

would be interested, on the contrary, in establishing conditions for the nonexistence of such cycles. In fact, in the first case a stationary oscillation is the desired goal while in the second case it is objectionable. Here the problem is the same but the purposes are different. The basic studies of this kind are obviously useful for further generalizations without which progress is impossible. Thus, for instance, instead of limiting oneself to the problem of establishing what happens in the presence of a given nonlinearity, one can foresee a much broader problem, namely, what kind of nonlinearity is to be introduced into the control system in order to predetermine its performance in a manner given in advance.

Such problems of synthesis are yet in a very early stage of their development.

D. H. Moore.[28] It is as if the author had written this paper for the writer's personal benefit, since it fits his present need so well. The writer's concern is to find a problem in nonlinear theory for a PhD thesis. There is a great deal of literature research to accomplish and the writer will be greatly assisted by the author's list of principal papers with comments on their particular values. It is heartening to see this important work so neatly presented and to find that the frontier is so nearly within reach.

G. J. Murphy.[29] This excellent and timely résumé should be of great value to the scientist who has been actively engaged in the analysis and synthesis of nonlinear control systems as well as to one who is about to enter this field. The thoroughness of the search which resulted in the publication of this paper is commendable, and the author's comments and suggestions concerning the relative values of the listed items can well result in the saving of much time for the reader.

However, although the procedure suggested in Section VIII of the paper for familiarization with published techniques is to be recommended to the research scientist, it is likely to be too time-consuming for the average engineer interested in acquiring a reasonable familiarity with practical methods of treating frequently encountered nonlinearities. For the latter group a more reasonable (and yet very profitable) approach is to study thoroughly Chapters 10 and 11 of Truxal's book[30] and then McDonald's work.[31, 32] Completion of such a program will provide the desired familiarity with practical approaches in a minimum of time and can always serve as a good basis from which to proceed to further study as desired.

L. M. Vallese.[33] The author gives a clear and informative picture of the history and status of nonlinear theory. His competent critical appraisal of the various methods of analysis and of the results so far obtained should encourage those who are not yet familiar with nonlinear theory to undertake its study with confidence.

There are various reasons for the scarce diffusion of such studies among engineers; the lack of a unified theory, the imperfection of the results, and the difficulties of application to synthesis problems are perhaps the most significant ones.

The theory has been developed by people with vastly different interests, such as astronomers, pure and applied mathematicians, mechanical, electronic, control, acoustical engineers, and so on. As a result the reader must digest papers with unfamiliar terminology, and then find out whether they may be applied to his own problems. In addition, while the perfection of linear theory has permitted its rationalization into "symbolic" or "operational" methods, which do not require *thinking*, the approximate nature of the procedures and results of nonlinear analysis make it imperative to control continuously the physical significance and the mathematical correctness of the various steps. Finally, the techniques of application to synthesis problems are quite different for linear and nonlinear cases. This is due to the fact that only in few instances nonlinear elements may be synthesized, while more often they constitute fixed components of the design. As a result, rarely one finds ready available analyses, and more often a direct study is necessary in order to derive the conditions of optimum design.

Ernst Weber.[34] The author is to be congratulated on the thoroughness with which he has explored the history of nonlinear control-system theory and the comprehensive enumeration of the methods of analysis.

This contribution will be very helpful to all who are interested in the existing literature on nonlinear analysis. In fact, the writer has direct use for it in connection with his own work in nonlinear-circuit theory.

Author's Closure

The author is most appreciative of the favorable comment by Professors Chin, Graham, Horgan, Ku, Kuba, Murphy, Vallese, and Weber and by Messrs. Black, Clark, Lanzkron, and Moore on the general values and usefulnesses of his paper. Mr. Bellman's note of recent developments which render dynamic programming a powerful method of nonlinear system analysis, Professor Fett's remark of combinations and particular applications of some of the methods listed in Section VI, and Mr. Forbes' statement of the recent commercial availability of digital differential analyzers complement, very pertinently, certain items in the paper.

Professor Minorsky's interesting discussion reaffirms and substantiates the author's own thesis, namely, that an accurate and complete insight to the behavior of nonlinear systems can be gained only by a thorough study and analysis of the actual nonlinear systems and associated solution of the corresponding nonlinear differential equations of performance; and not by approximating them by linear systems—although this last procedure may provide useful information pertinent to certain limited conditions of operation. It is precisely on the basis of this thesis, stressed both by him and by Professor Minorsky, that the author must differ with the remark in the last paragraph of Professor Murphy's discussion. Thus Chapters 10 and 11 of Truxal's book[30] give, respectively, a good account of the so-called describing-function technique and an introduction to some topological aspects of nonlinear analysis; and MacDonald's two papers[31, 32] afford an introduction (they have been followed by some 15 papers extending his initial work) to multiple-mode optimization of relay-type servomechanisms. But the whole of the knowledge encompassed in these just-mentioned sources is only a minute fraction of the existing total of nonlinear-system analysis which is available for use in control-system design and synthesis; and it is to the engineer who desires to make serious use of this available total, rather than merely learn fragments of it, that this paper is addressed.

Finally, the author extends his sincere thanks to all of the dis-

[28] Santa Monica, Calif.

[29] Assistant Professor of Electrical Engineering, University of Minnesota, Minneapolis, Minn.

[30] "Automatic Feedback Control-System Synthesis," by J. G. Truxal, McGraw-Hill Book Company, Inc., New York, N. Y., 1955.

[31] "Nonlinear Techniques for Improving Servo Performance," by D. McDonald, Proceedings of the National Electronics Conference, vol. 6, 1950, pp. 400–421.

[32] "Multiple Mode Operation of Servomechanisms," by D. McDonald, *Review of Scientific Instruments*, vol. 23, January, 1952, pp. 22–30.

[33] Associate Professor of Electrical Engineering, Polytechnic Institute of Brooklyn, Brooklyn, N. Y.

[34] Head, Department of Electrical Engineering, Polytechnic Institute of Brooklyn, Brooklyn, N. Y.

cussers for the complimentary remarks and helpful comments encompassed in their discussion as a whole.

In conclusion, the author would direct attention to the recent excellent survey paper by Clauser[35] (to be considered as an additional item relative to Section III on Orientation); would stress that the Bibliography comprises books, survey articles, and special treatments of broad phases of theory as discussed in the context, and is not intended to include papers on specific problems or particular analyses; and would note that a separate, very exhaustive bibliography of some 400 such papers is available (from the Engineering Experiment Station, University of Wisconsin, Madison, Wis.) on request, as mentioned in the last paragraph of the paper.

[35] "The Behavior of Nonlinear Systems," by F. H. Clauser, *Journal of the Aeronautical Sciences*, vol. 23, 1956, pp. 411–434.

Electrohydraulic Servomechanism With an Ultrahigh-Frequency Response[1]

By D. P. ECKMAN,[2] C. K. TAFT,[3] AND R. H. SCHUMAN,[4] CLEVELAND, OHIO

This paper presents an analysis of a positional servomechanism of very high performance. This control employs an electronic error transducer which actuates a pilot valve by means of an amplifier and a torque motor to position a hydraulic cylinder. The control was to have a frequency response whose amplitude ratio was nearly 1 to 200 cycles per sec (cps) at an amplitude of 0.001 in. with 100-lb dry friction and a load mass of 200 lb. These specifications were exceeded by the control described herein. The optimum open-loop gain and the closed-loop frequency response are determined by linearizing the system equations and using Laplace transform methods. The system also was analyzed by solving the nonlinear equations on an electronic analog computer to determine optimum gain, transient response, and frequency response. A comparison of the results indicates that for input signals which cause the control to operate outside the region in which the linearizing assumptions apply, the linear analysis still gives results which agree with those of the nonlinear analysis within a factor of three.

Nomenclature

The following nomenclature is used in the paper:

a = equilibrium value of valve spool nozzle opening, in.
A_c = area of large side of piston, sq in.
A_v = area of large end of valve spool, sq in.
b = equilibrium opening of nozzle 2 in valve, in.
B = isothermal bulk modulus of oil, psi
c = displacement of piston from equilibrium position, in.
C = transformed cylinder displacement, volts
$\dot{c}$ = dimensionless cylinder velocity
D_1 = diameter of spool nozzle 1, in.
D_2 = diameter of valve nozzle 2, in.
D_v = valve-spool diameter at ports, in.
e' = error signal from pickup, volts
e = error $= c - r$, in.
E' = transformed error signal, volts
E = transformed error $= C - R$, volts
f = cross cylinder force, lb
f_c = cylinder compressibility frequency, cycles per sec (cps)
f_f = friction force on piston, lb
f_t = constant force on piston, lb
F_f = transformed piston friction force, volts
F_t = transformed piston constant force, volts
$\bar{f}$ = dimensionless gross cylinder-output force
K = flow coefficient, $\text{in}^3/\text{sec lb}^{1/2}$
K_1 = change of flow rate through nozzle 1 in spool with valve opening, in^2/sec
K_2 = change of flow rate through nozzle 1 in spool with p_c, $\text{in}^5/\text{lb sec}$
K_3 = change of flow rate through nozzle 1 in spool due to change in supply pressure, $\text{in}^5/\text{lb sec}$
K_4 = change of flow rate through nozzle 2 in valve due to torque-motor displacement, in^2/sec
K_5 = change of flow rate through nozzle 2 in valve due to change in p_c, $\text{in}^5/\text{lb sec}$
K_6 = change of cylinder force with cylinder velocity at constant valve opening, lb sec/in.
K_7 = change of cylinder force with valve opening at constant cylinder velocity, lb/in.
K_8 = change in flow rate to cylinder with cylinder pressure, $\text{in}^5/\text{lb sec}$
K_9 = change in flow rate to cylinder with valve opening, in^2/sec
K_A = torque-motor-amplifier gain, $\frac{\Delta \text{ ma}}{\text{volt}}$
K_d = torque motor-differential transformer gain, volts/in.
K_f = torque motor-feedback amplifier gain, volts/volt
K_p = pickup-tube gain, volts/in.
K_t = torque-motor gain, in/Δ ma
K_v = pilot-valve gain, in/in.
M_L = load mass, $\text{lb sec}^2/\text{in}$.
M_v = valve-spool mass, $\text{lb sec}^2/\text{in}$.
p_c = oil pressure on large side of piston, psig
p_{c0} = equilibrium value of p_c, psig
p_c' = deviation of p_c from p_{c0}, psig
P_c = transformed cylinder pressure, volts
p_s = supply pressure, psig
p_s' = deviation of p_s from equilibrium value, psi
p_v = oil pressure on large side of valve spool, psig
p_v' = deviation of p_v from equilibrium value, psig
Q_c = oil-flow rate to cylinder, in^3/sec
Q_{p1} = oil-flow rate through nozzle 1 in valve spool, in^3/sec
Q_{p10} = equilibrium value of Q_{p1}, in^3/sec
Q_{p2} = oil flow through nozzle 2 in valve block, in^3/sec
Q_{p20} = equilibrium value of Q_{p2}, in^3/sec
Q_{v1} = oil-flow rate through valve-supply port, in^3/sec
Q_{v10} = equilibrium value of Q_{v1}, in^3/sec
Q_{v2} = oil-flow rate through valve drain port, in^3/sec
Q_{v20} = equilibrium value of Q_{v2}, in^3/sec
r = reference input, in.
R = transformed reference input, volts
s = Laplace operator
S = transformed Laplace operator
x_0 = equilibrium piston position measured from position when minimum clearance exists on the valve side of the cylinder, in.

[1] Portions of this paper were taken from Mr. Taft's thesis, "An Analysis of an Electrohydraulic Positional Servomechanism," submitted in partial fulfillment of the Degree of Master of Instrumentation Engineering at Case Institute of Technology, Cleveland, Ohio. Information concerning the complete thesis may be obtained from the Librarian, Case Institute of Technology.

[2] Professor of Mechanical Engineering, Case Institute of Technology.

[3] Research Engineer, Warner and Swasey Company. Assoc. Mem.

[4] Engineering Physicist, Warner and Swasey Company.

Contributed by the Instruments and Regulators Division of THE AMERICAN SOCIETY OF MECHANICAL ENGINEERS and presented at the ASME-AIEE Conference on Nonlinear Control Systems, Princeton, N. J., March 26–28, 1956.

 Manuscript received at ASME Headquarters, January 10, 1956. Paper No. 56—IRD-8.

T_c = cylinder compressibility time constant, sec
T_c' = cylinder velocity time constant, sec
T_t = torque-motor time constant, sec
T_v = pilot-valve time constant, sec
u = pilot-valve underlap, in.
V = volume of oil in large side of cylinder and in line from valve to cylinder, cu in.
V_0 = minimum value of V, cu in.
x = valve displacement, in.
$\hat{x}$ = dimensionless valve displacement
X = transformed valve motion, volts
x_m = maximum valve motion, in.
y = torque-motor displacement, in.
y_m = maximum torque-motor motion, in.
Y = transformed torque-motor displacement, volts
α = pickup-linkage ratio
ζ_c = cylinder damping ratio
ζ_t = torque-motor damping ratio
ζ_v = valve damping ratio
ψ = generalized equilibrium cylinder position = $s_0 + \frac{V_0}{A_c}$, in.
Ψ = transformed equilibrium position, volts
ρ = oil mass density, lb sec^2/in^4
τ = transformed computing time, sec.

Introduction

The trend today is toward increased automation to speed up production processes. Many processes require high-power and high-performance positional controls in order to be automatized and this is especially true in the field of machine tools. Hydraulic controls employing hydraulic cylinders and valves can be designed to deliver large amounts of power with good dynamic performance. However, a mechanical input is often required, such as the displacement of a cam follower riding on a cam. Electronic components may be used along with the hydraulic cylinder and valve to permit more flexible and compact arrangement of error-sensing devices, remote location of error-sensing devices, dynamic compensation of the system transfer characteristics, easily adjustable loop gain, and electrical reference inputs. Thus a more versatile control can be designed using both electronic and hydraulic components.

The electrohydraulic positional servomechanism that is to be analyzed uses electronic components in the low-power input section and hydraulic components in the high-power output section.

In this paper will be described the methods of linearization and linear analysis and the use of the electronic analog computer in solving the nonlinear equations for the dynamic performance of the control. The region of validity of the linear analysis is indicated in the comparison of the results obtained by the two methods of analysis.

The design of high-performance controls has brought forth many problems. Resonances in the open-loop frequency response due to load mass and oil compressibility, along with resonances of the control-mounting structure, can limit the response of high-performance positional controls. Any resonant frequencies of the elements in the control loop which are near or below the desired cutoff frequency can cause this type of servomechanism to be unstable except at very low open-loop gains. When the open-loop gain is low, good closed-loop performance is impossible at high frequencies. The problems involved in control stabilization with mount structural resonances and oil compressibility resonances are the same from a control standpoint.

Linear Analysis

The transfer function of each element in the control will be derived from the differential equations. The open-loop frequency-response function will be obtained from the transfer functions and will be used to determine the optimum open-loop gain.

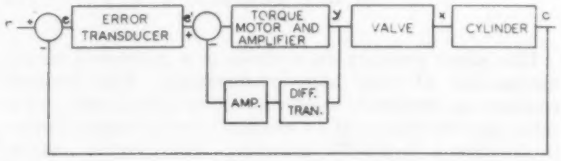


Fig. 1 Block Diagram of Electrohydraulic Servomechanism

The control is shown in block diagram form in Fig. 1. The error transducer is mounted on the piston of the hydraulic cylinder and contacts a template. When an error exists between the template and piston positions, the transducer output voltage changes. This voltage change is the input to a torque-motor amplifier whose output is a difference in current to the coils of a torque motor. This differential current displaces the torque-motor armature. The motion of the armature is detected by a differential transformer whose output is a modulated voltage proportional to armature displacement. This voltage is demodulated, amplified, and subtracted from the transducer output-voltage change to correct for armature-position errors. The torque-motor armature controls the opening of a nozzle which is in series with a nozzle in the valve spool whose opening depends on spool position. A motion of the torque-motor armature causes an unbalance of the pressure-area forces on the valve spool imparting motion to it in the direction which restores the pressure balance. The pilot spool controls the flow of oil to the cylinder imparting a velocity to the piston to correct the error which caused the transducer signal.

Error Transducer. The error transducer is a triode electronic tube with a movable anode (Radio Corporation of America No. 5734). Displacement of the anode causes a proportional variation in anode voltage. The tube is mounted with a linkage so that the anode motion is restricted and is a fraction of the control error.

The moving element of the tube has a natural frequency of 12,000 cps. The linkage was designed so that its resonant frequency was above 500 cps.

The equations for each element of the system are given in Appendix 1.

Torque Motor and Amplifier. The torque motor is a device which converts electrical signal into a displacement. The armature is a flat plate that has two rods attached to its center which are anchored at their ends. The axis of these rods is the axis of rotation of the plate. Mounted around the armature on opposite sides of the torsion rods are two coils. This assembly is located between two magnets, the poles of which are separated by a small gap from each end of the armature. When more current flows through one coil than the other (a differential current exists in the coils) the resulting net flux in the armature supports the flux flowing between the magnets on one end of the armature and opposes the flux on the other end. This causes a torque on the armature in the direction of the maximum flux which twists the torsion rods displacing the armature.

The torque motor used here was a Midwestern Geophysical Laboratory Model 9.

Referring to Fig. 2, the torque-motor armature operates against two open nozzles which in turn control the oil flow from the pilot-

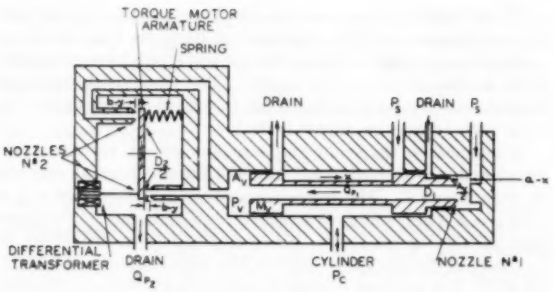


FIG. 2 SCHEMATIC DIAGRAM OF PILOT VALVE

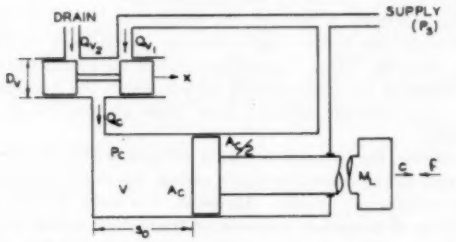


FIG. 3 SCHEMATIC DIAGRAM OF CYLINDER AND VALVE

valve nozzle and therefore position the pilot-valve spool as described in the next section.

Pilot Valve. The pilot valve is shown schematically in Fig. 2, and the operation is as follows: When the torque-motor displacement y increases, the area of opening of the outlet nozzles decreases. Thus the control pressure p_v increases, exerting a net force on the valve spool causing it to move to the right. This increases valve displacement x. Increasing valve displacement decreases the opening of nozzle 1. Since supply pressure is assumed constant, increasing the pilot-valve displacement causes the control pressure to decrease to its equilibrium value. The area of the supply-pressure side of the spool is equal to one half the area of the other side so that the equilibrium value of the control pressure is one half supply pressure. This type of arrangement permits an input displacement at low power level to position the valve spool with good accuracy at high frequencies.

The spool is constructed so that there is negligible pressure drop across any bearing surface. Binding of the valve spool may be caused by variations in pressure around the circumference of the spool caused by nonconstant radial clearances between the spool and the valve block. The bearing lengths are made short by grinding the diameter of all but a short length of the spool lands slightly under the diameter of the bearing surface. The ends of the bearing surfaces are grooved and joined by drilled holes so that essentially no pressure drop exists across these bearing surfaces (1).[5] This reduces spool friction to a very small value.

The flow through the nozzles will be assumed to be the same as that through a sharp-edged orifice of the same area. It has been found that for a discharge coefficient of 0.61 this assumption is correct for many different nozzle configurations. This is not correct for very small flows or for nozzles whose area is greater than about one tenth the area upstream (2, 3). The nozzles are operated at openings of less than one quarter their diameter, so that the nozzle area is equal to the circumference times the nozzle opening.

[5] Numbers in parentheses refer to the Bibliography at the end of the paper.

Leakage by the spool lands and compressibility of the entrapped oil will be neglected since both have little effect here.

Cylinder. The cylinder used is a single-acting type with constant back pressure, as indicated in Fig. 3. The area of the constant-pressure side is equal to one half that of the variable-pressure side. Flow to the cylinder is controlled by a zero-lapped valve with no radial clearance which ideally has no leakage when it is centered in the valve block. This type of pilot is impossible to manufacture with no clearance; however, small clearances have a small effect on performance, as will be shown.

The force-velocity curves for the pilot valve and cylinder combination are described in Appendix 2.

Open-Loop Linear Analysis. The complete system equations in linear form as shown in Appendix 1 are as follows:

(*a*) Error transducer open loop

$$\frac{e'}{e} = \frac{K_p}{\alpha} \quad \ldots\ldots [1a]$$

(*b*) Torque motor and amplifier with loop closed around it

$$\frac{y}{e'} = \frac{K_A K_t}{T_t^2 s^2 + 2\zeta_t T_t s + 2} \quad \ldots\ldots [3]$$

(*c*) Pilot valve

$$\frac{x}{y} = \frac{K_v}{T_v^2 s^2 + 2\zeta_v T_v s + 1} \quad \ldots\ldots [9]$$

(*d*) Cylinder

$$\frac{c}{x} = \frac{1}{T_c' s (T_c^2 s^2 + 2\,\zeta_c T_c s + 1)} \quad \ldots\ldots [20]$$

Combining these the open-loop transfer function is

$$\frac{c}{e} = \frac{K_p K_A K_t K_v}{\alpha T_c' s\,(T_c^2 s^2 + 2\zeta_c T_c s + 1)} \times (T_v^2 s^2 + 2\zeta_v T_v s + 1)(T_t^2 s^2 + 2\zeta_t T_t s + 2) \quad \ldots [21]$$

The value of K_A must be determined for optimum closed-loop response. As an approximate criterion K_A should be adjusted so that

$$\left|\frac{c}{e}\right| = 1$$

at 135-deg phase lag. As long as the value of

$$\left|\frac{c}{e}\right| < 0.7$$

at 180-deg phase lag this criterion may provide satisfactory response.

The effect of the oil compressibility and load mass is to cause a peak in the cylinder-amplitude ratio at a "resonant" frequency. As the equilibrium piston position moves toward larger volume (to the right in Fig. 3) the resonant peak occurs at lower frequencies. At this resonant frequency the phase lag between valve position x and piston position c is 180 deg. Thus as cylinder stroke is increased, the frequency at which the control has 180-deg phase lag is decreased. As this frequency decreases the value of

$$\left|\frac{c}{eK_A}\right|$$

at 180 deg, phase lag increases (see Figs. 4 and 5). Thus the open-loop gain must be reduced as the oil compressibility frequency decreases. In this system, lower gain decreases the

cutoff frequency of the closed-loop frequency response and decreases the rate at which the piston corrects position errors. Using a Nichols chart it can be shown that for a given system, decreasing open-loop gain causes the cutoff frequency to be lower. Therefore, for high performance, it is desirable to have the oil compressibility frequency as high as possible.

The resulting open-loop response with $K_A = 4.7\ \Delta$ ma/volt is plotted in Figs. 4 and 5.

Closed-Loop Frequency Response. The closed-loop transfer function is

$$\frac{c}{r} = \frac{(K_p K_A K_t K_v)/\alpha}{(K_p K_A K_t K_v)/\alpha + T_s' s(T_e^2 s^2 + 2\zeta_e T_e s + 1) \times (T_v^2 s^2 + 2\zeta_v T_v s + 1)(T_t^2 s^2 + 2\zeta_t T_t s + 2)} \quad [23]$$

With actual data for the system (Appendix 1) the function becomes

$$\frac{c}{r} = \frac{1}{1.15 \times 10^{-23} s^7 + 9.96 \times 10^{-21} s^6 + 6.3 \times 10^{-17} s^5 + 2.3 \times 10^{-13} s^4 + 6.59 \times 10^{-10} s^3 + 1.09 \times 10^{-6} s^2 + 1.4 \times 10^{-3} s + 1} \quad [23a]$$

The closed-loop frequency response is calculated and plotted in Figs. 6 and 7, where it is compared to the closed-loop response obtained by the computer.

Nonlinear Analog-Computer Analysis

System Equations. The system equations given in Appendix 1 were set up on a George A. Philbrick Researches, Inc., analog computer.

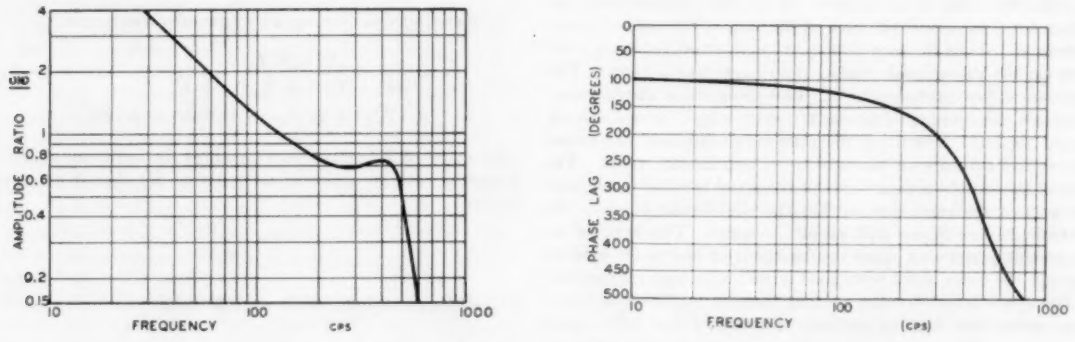
Closed-loop response to a step input was recorded under varying conditions. A large step input of 0.004 in. was chosen because in many applications of this control it will have to respond with fidelity to large step inputs. The step input was large enough to represent a typical input but small enough so that the error-detecting device did not limit.

A closed-loop frequency response was run at various amplitudes and was plotted along with the linear-frequency response in Figs. 6 and 7. The computer analysis indicated that a gain of $K_A = 2\ \Delta$ ma/volt gave satisfactory performance. The phase lags at varying frequency correspond closely to the linear response, but the computer response was less oscillatory at the gain of $K_A = 2\ \Delta$ ma/volt and the amplitude ratio fell to 0.82 before rising to a peak of 1.25 at 320 cps for large amplitudes, and a peak of 1.95 at 350 cps for small amplitudes.

This discrepancy between the results of the two methods of analysis is due to the large reference inputs used in the nonlinear analysis. These large inputs caused the cylinder to operate in regions of the force-velocity curves where the dynamic characteristics were far different from those assumed in the linear analysis.

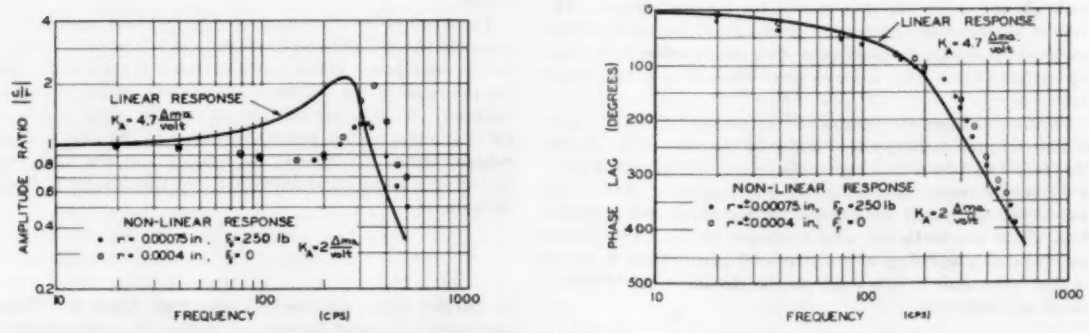
However, the two sets of results do agree within limits. Since the gain in the control is easily varied, the linear analysis gave quite satisfactory results. However, if a more accurate determination of frequency response or transient response is desired at large amplitudes where linearization techniques do not apply, the analog-computer solution of the nonlinear equations is necessary.

The effect of a constant force f_t on the response is shown in Fig. 8. Increased constant force causes the rise time to increase.



$K_A = 4.7\ \Delta$ ma/volt

Figs. 4, 5 Linearized Open-Loop Frequency Response



Figs. 6, 7 Computed Closed-Loop Frequency Response

This is due to the limiting of spool motion as well as the effect on dynamic performance of the flattening of the cylinder characteristics. Referring to Fig. 9, it is seen that under large constant forces the maximum piston velocity decreases and the force-velocity curves flatten out and converge. This accounts for this slow response.

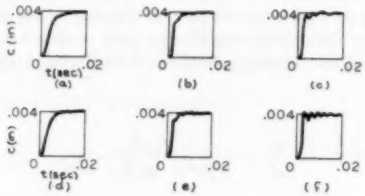
Final Test Results

The system described was built and tested at Case Institute of Technology. Because of various improvements made in the many control elements it was unfortunately not possible to test exactly the same system as was described in the mathematical and computer analysis. The actual response of the torque motor with the loop closed around it had indicated a lower natural frequency and higher damping ratio than was assumed for the linear analysis. The load mass was not as high but the cylinder-equilibrium position was much larger so that compressibility frequency was not radically altered in analysis and test. A summary of these conditions is given in Table 1.

The test results were obtained on a system mounted on a large and rigid I-beam structure fastened to a heavy concrete floor.

TABLE 1 COMPARISON OF TEST CONDITIONS

	Linear analysis	Nonlinear computer analysis	Actual test
Torque motor f_n, cps (closed loop)	470	470	350
Load mass, lb	200	200	20
Dry friction, lb	0	0 and 250	10
Cylinder equil. position, in.	0.5	0.5	1.25
Output amplitude, in.	...	0.0004 and 0.00075	0.001
Amplifier gain, Δ ma/volt	4.7	2.0	1.6
Loop gain, sec^{-1}	1400	600	500



K_A = 2 Δ ma/volt, y = 0.5 in.

(a) f_t = 2000 lb f_f = 250 lb
(b) f_t = 1000 lb f_f = 250 lb
(c) f_t = 0 f_f = 250 lb
(d) f_t = 2000 lb f_f = 0
(e) f_t = 1000 lb f_f = 0
(f) f_t = 0 f_f = 0

Fig. 8 Closed-Loop Transient Response From Nonlinear Analog-Computer Analysis

The hydraulic-power supply must provide reasonably smooth pressure but no accumulators were used. Reinforced rubber hose was used for power-supply lines.

The closed-loop response of the system is shown in Figs. 10 and 11. The response is quite stable although oscillatory. The frequency response indicates that the transient response (not shown here) would oscillate several cycles before equilibrium was reached.

Conclusions

The compressibility of the entrapped oil in the large side of the cylinder between the valve and the piston causes the cylinder to have an oscillatory response. If the resonant frequency of this oscillatory component is not well above the desired cutoff frequency of the control, it will introduce enough phase lag to lower the cutoff frequency. This becomes a problem when the desired cutoff frequency is above 150 cps using a hydraulic cylinder of less than 4-in. diam at strokes of more than 2 in. To increase the compressibility resonant frequency the cylinder diameter should be increased, its stroke decreased, and its load mass decreased.

The methods of analysis used here indicate that, for a control of this type, which employs a zero-lapped pilot and hydraulic cylinder, linearized dynamic analysis provides better than order-of-magnitude results when compared to the results from the nonlinear analysis. If accurate values of the open-loop gain and transient response to large input signals are desired, the nonlinear analysis employing an analog computer must be used.

At small equilibrium-cylinder positions the system-transient

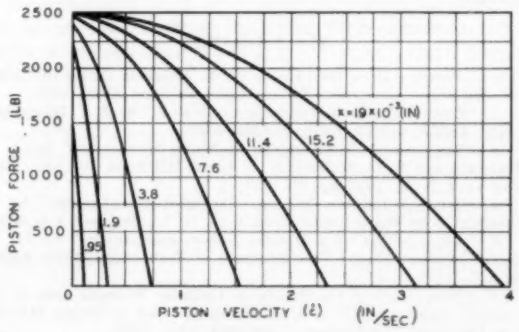
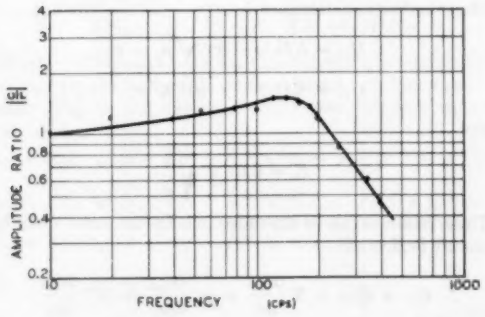
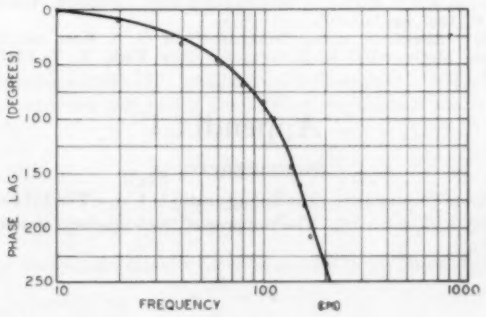


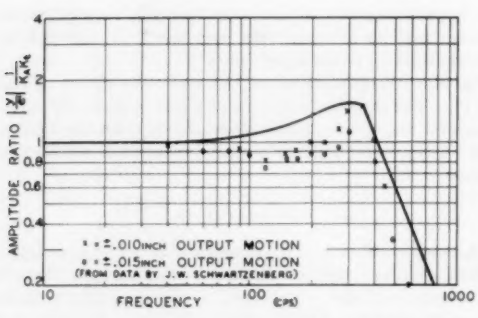
Fig. 9 Valve-Cylinder Force-Velocity Curves Corrected for Valve Clearances

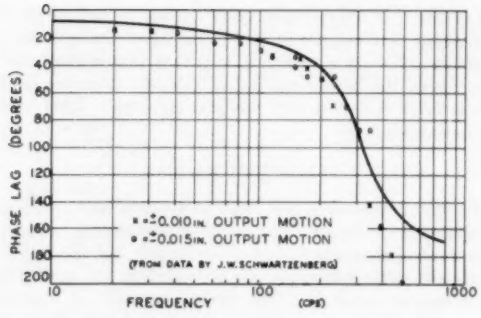




K_A = 1.6 Δ ma/volt

Figs. 10, 11 System Closed-Loop Frequency-Response Test Results





FIGS. 12, 13 TORQUE MOTOR AND AMPLIFIER OPEN-LOOP FREQUENCY RESPONSE

performance is good and it has a frequency response which is nearly flat to 400 cps.

Dry friction of the type which is a constant force opposing piston velocity, seems to be desirable if the system is to be operated at long cylinder strokes with large mass loads.

ACKNOWLEDGMENTS

The authors wish to acknowledge the support of this work at Case Institute of Technology by the Warner & Swasey Company.

Significant contributions were made by a number of the graduate students in Instrumentation. Mr. George M. Lance investigated the pilot-valve systems. Mr. Jack W. Schwartzenberg pioneered the amplifier and torque motor, which was further carried on by Mr. Steve Hemlar and Mr. Lee Gallaher. Mr. William F. Ortman selected the error transducer.

BIBLIOGRAPHY

1 "Pressure Bind Relief," by D. P. Eckman, Notes in Course ME 273, Case Institute of Technology, Cleveland, Ohio, 1955.

2 "Contributions to Hydraulic Control, 3," by J. F. Blackburn, Trans. ASME, vol. 75, 1953, pp. 1163–1170.

3 "Engineering Applications of Fluid Mechanics," by J. C. Hunsaker and B. G. Rightmire, McGraw-Hill Book Company, Inc., New York, N. Y., 1947, pp. 157–158.

4 "Servomechanisms and Regulating System Design," by H. Chestnut, John Wiley & Sons, New York, N. Y., 1951, vol. 1, p. 331.

5 "An Analysis of a Hydraulic Servomechanism," by G. M. Lance, Master's thesis, Case Institute of Technology, 1954, Appendix 1.

6 "Frequency of Oscillation of Cylinder Systems Due to the Compressibility of Oil," by D. P. Eckman, Notes in Course ME 273, Case Institute of Technology, Cleveland, Ohio, 1955.

7 Reference (4), pp. 311–314.

8 "Principles of Servomechanisms," by G. S. Brown and D. P. Campbell, John Wiley & Sons, New York, N. Y., 1948, p. 107.

9 "Theory of Servomechanisms," by H. M. James, N. B. Nichols, and R. S. Phillips, McGraw-Hill Book Company, New York, N. Y., 1947, pp. 179–185.

10 "Electronic Analog Computers," by G. A. Korn and T. M. Korn, McGraw-Hill Book Company, New York, N. Y. 1947, pp. 157–158.

Appendix 1

SYSTEM EQUATIONS

Error Transducer. The linkage ratio is $1/\alpha$. The tube sensitivity is K_p volts/in. The response of the transducer is assumed to be

$$e' = \frac{K_p}{\alpha} e \qquad [1]$$

where $e = c - r$ when the system loop is closed.

Torque Motor and Amplifier. The torque motor is a Midwestern Geophysical Laboratory Model 9. It is driven by an amplifier with high current feedback so that the droop in frequency response caused by the inductance and resistance of the torque-motor coils is minimized. The experimental torque-motor response is shown in Figs. 12 and 13. It is nonlinear; however, its response can be approximated by a second-order linear differential equation as shown by the solid curve, assuming 9-deg phase lag at zero frequency.

The torque-motor and amplifier transfer function (linear part) can then be written

$$\frac{y}{e'} = \frac{K_A K_t}{T_t^2 s^2 + 2\zeta_t T_t s + 1} \qquad [2]$$

To improve torque-motor response a loop was closed around it and loop gain was adjusted to 1 at 135-deg phase lag. It was assumed the differential transformer used to detect torque-motor position and the feedback amplifier contributed negligible phase lag.

Thus

$$\frac{y}{e'} = \frac{K_A K_t}{T_t^2 s^2 + 2\zeta_t T_t s + 2} \qquad [3]$$

since

$$K_A K_t K_f K_d = 1$$

Pilot Valve. The nozzles are operated at openings of less than one quarter their diameters so that the nozzle area is equal to the circumference times the nozzle opening. Line losses between nozzles 1 and 2 are very small and will be neglected. Drain pressure will be assumed to be equal to atmospheric pressure, since line losses between the drain and the sump will be neglected in this analysis.

Thus the flow equations are

$$Q_{p1} = KD_1(a - x)\sqrt{p_s - p_v} \qquad [4]$$

$$Q_{p2} = KD_2(b - y)\sqrt{p_v} \qquad [5]$$

where

$$K = 0.61\,\pi\sqrt{\frac{2}{\rho}}$$

These relations can be expanded in a Taylor series and if higher terms are neglected

$$Q_{p1} = Q_{p10} + X\frac{\partial Q_{p1}}{\partial x} + p_s'\frac{\partial Q_{p1}}{\partial p_s} + p_v'\frac{\partial Q_{p1}}{\partial p_v}$$

where p_s' and p_v' are deviations of pressure about the mean values when $Q_{p1} = Q_{p10}$. These partial derivatives are variable

but for small values of x, p_s', p_v' they can be assumed constant and the equations can be linearized about some point.

In this analysis these equations will be linearized about the point where

$$\left.\begin{aligned} x = 0,\ p_v = p_s/2 \quad &\text{and} \quad Q_{p1} = Q_{p2} \\ Q_{p1} = Q_{p10} - x\,K_1 - p_v'K_2 + p_s'K_3 \end{aligned}\right\} \quad \ldots\ldots [4a]$$

Similarly

$$Q_{p2} = Q_{p20} - yK_4 + p_v'K_5 \quad \ldots\ldots [5a]$$

where

$$K_1 = -\frac{\partial Q_{p1}}{\partial x} = \frac{Q_{p1}}{a-x} \qquad K_4 = -\frac{\partial Q_{p2}}{\partial y} = \frac{Q_{p2}}{b-y}$$

$$K_2 = -\frac{\partial Q_{p1}}{\partial p_v} = \frac{Q_{p1}}{2(p_s - p_v)} \qquad K_5 = \frac{\partial Q_{p2}}{\partial p_v} = \frac{Q_{p2}}{2p_v}$$

$$K_3 = \frac{\partial Q_{p1}}{\partial p_s} = -K_2$$

Neglecting spool leakage and compressibility of the entrapped oil, both of which have little effect here, the flow-velocity relation can be written

$$\left.\begin{aligned} A_v\dot{x} &= Q_{p1} - Q_{p2} \\ &= A_v\dot{x} = -xK_1 - p_v'K_2 + p_s'K_3 + yK_4 - p_v'K_5 \end{aligned}\right\} \ldots [6]$$

Summing forces on the valve spool, neglecting viscous damping, friction, the increase in spool inertia due to moving oil and flow forces, yields

$$M_v\ddot{x} = A_v\left(p_v' - \frac{p_s'}{2}\right) \quad \ldots\ldots [7]$$

Equations [6] and [7] can be combined

$$M_v\ddot{x} = \frac{A_v}{K_2 + K_5}\left[-K_1x + K_3p_s' + K_4y - A_v\dot{x}\right] - p_s'\frac{A_v}{2} \quad \ldots [8]$$

but

$$\frac{A_vK_3}{K_2 + K_5} - \frac{A_v}{2} = 0$$

Canceling and rearranging yields

$$\frac{x}{y} = \frac{K_v}{T_v^2s^2 + 2\zeta_vT_vs + 1} \quad \ldots\ldots [9]$$

where

$$T_v = \sqrt{\frac{M_v(K_2 + K_5)}{K_1A_v}}, \quad \zeta_v = \frac{A_v}{2K_1}\sqrt{\frac{K_1A_v}{M_v(K_2 + K_5)}}, \quad K_v = \frac{K_4}{K_1}$$

Cylinder. The valve has a slight radial clearance and underlap. Referring to the schematic representation of the valve and cylinder (Fig. 3) and assuming orifice flow, the flow equations can be written

$$\left.\begin{aligned} Q_{c1} &= KD_v\sqrt{(x+u)^2 + \overline{\Delta r}^2}\sqrt{p_s - p_c} \\ Q_{c2} &= KD_v\sqrt{(u-x)^2 + \overline{\Delta r}^2}\sqrt{p_c} \end{aligned}\right\} \begin{cases} u \geq x \geq 0 \\ -u \leq x \leq 0 \end{cases}$$

$$\left.\begin{aligned} Q_{c1} &= KD_v\sqrt{(x+u)^2 + \overline{\Delta r}^2}\sqrt{p_s - p_c} \\ Q_{c2} &= KD_v\,\Delta r\sqrt{p_c} \end{aligned}\right\} \quad x \geq u$$

$$\left.\begin{aligned} Q_{c1} &= KD_v\,\Delta r\sqrt{p_s - p_c} \\ Q_{c2} &= KD_v\sqrt{(u-x)^2 + \overline{\Delta r}^2}\sqrt{p_c} \end{aligned}\right\} \quad x \leq u$$

$$\ldots\ldots [10a]$$

where

$$K = 0.61\,\pi\sqrt{\frac{2}{\rho}}$$

u = underlap

Δr = radial clearance between valve spool and block

In the computer analysis these equations were simplified by assuming that $u = 0$ and $\Delta r = 0$. For the linear analysis, Equations [10] were linearized in the region of valve opening: $0.006 \leq x \leq 0.008$, and $p_c = p_s/2$, in the same manner as the pilot-valve flow equations

$$\left.\begin{aligned} Q_{v1} &= K_9x - K_8p_c' + Q_{v10} \\ Q_{v2} &\cong Q_{v20} \end{aligned}\right\} \quad \ldots\ldots [10b]$$

where in the region of $x \geq u$

$$K_9 = \left.\frac{\partial Q_{v1}}{\partial x}\right| = \frac{Q_{v1}(x+u)}{(x+u)^2 + \overline{\Delta r}^2}, \quad K_8 = -\frac{\partial Q_{v1}}{\partial p_c'} = \frac{Q_{v1}}{p_s - p_c}$$

Now, assuming isothermal compressibility of the fluid on the valve side of the piston, the relation between flow and velocity can be written, neglecting line losses between the valve and cylinder

$$A_c\dot{c} = Q_{v1} - Q_{v2} - \frac{V}{B}\dot{p}_c \quad \ldots\ldots [11]$$

where B = isothermal bulk modulus of the hydraulic oil.

By definition $V = (s_0 + c)A_c + V_0$. In this analysis, dynamic variations about an equilibrium piston position s_0 will be examined where $s_0 \gg c$. Thus

$$V \cong s_0A_c + V_0 \equiv \psi A_c$$

where ψ is the generalized equilibrium cylinder position.

A force balance on the piston yields

$$M_L\ddot{c} = \left(p_c' - \frac{p_s}{2}\right)A_c - f_f - f_t + p_{s0}A_c \quad \ldots\ldots [12]$$

For static balance

$$p_{c0}A_c - p_s\frac{A_s}{2} - f_t = 0$$

In the linear analysis it is assumed that friction force f_f is zero.

In the nonlinear computer analysis friction force f_f is assumed equal to

$$\frac{\dot{c}}{|\dot{c}|}|f_f|$$

Equations [10b], [11], and [12] can be combined. Neglecting constant terms and taking the Laplace transform of the result yields

$$M_Ls^2c = \left[\frac{K_9x - A_cs\bar{c}}{K_8 + \frac{\psi}{B}A_cs}\right]A_s \quad \ldots\ldots [13]$$

Thus the cylinder transfer function as used in the linear analysis is

$$\frac{c}{X} = \frac{1}{T_c'^2s(T_c^2s^2 + 2\zeta_cT_cs + 1)} \quad \ldots\ldots [14]$$

where

$$T_e' = \frac{A_e}{K_v}, \quad T_e = \sqrt{\frac{M_L \Psi}{B A_e}}, \quad \zeta_e = \sqrt{\frac{M_L K_v^2 B}{4 A_e^3 \Psi}}$$

Equations [10a] and [11] can be combined to yield

$$\left.\begin{aligned} x \geq 0, \quad A_e \dot{c} &= K D_v x \sqrt{p_s - p_c} - \frac{V}{B} \dot{p}_c \\ x \leq 0, \quad A_e \dot{c} &= K D_v x \sqrt{p_c} - \frac{V}{B} \dot{p}_c \end{aligned}\right\} \quad [15]$$

Equations [12] and [15] were used to describe cylinder dynamic response in the nonlinear computer analysis.

Numerical Values

In the linear analysis the following were the values of the constants

$$\begin{aligned} \frac{K_F}{\alpha} &= 1830 \text{ volts/in.} \\ K_A &= \text{determined by dynamic analysis} \\ K_i &= 7.5 \times 10^{-4} \text{ in}/\Delta\text{ma} \\ \zeta_i &= 0.33 \\ T_i &= 4.82 \times 10^{-4} \text{ sec} \\ T_v &= 2.94 \times 10^{-4} \text{ sec} \\ \zeta_v &= 0.673 \\ K_v &= 1.555 \\ T_e' &= 7 \times 10^{-3} \text{ sec} \\ T_e &= 2.87 \times 10^{-4} \text{ sec} \\ \zeta_e &= 0.386 \end{aligned}$$

The open-loop frequency response is plotted in Figs. 4 and 5. The values for the nonlinear computer analysis were the same except in the case of the cylinder.

The following equations were used to describe the cylinder response:

For $x \geq 0$, $sc = 14.9 \times \sqrt{400 - p_c} - \frac{4\Psi s}{10^3} p_c' \quad [15a]$

For $x \leq 0$, $sc = 14.9 \times \sqrt{p_c} - \frac{4\Psi s}{10^3} p_c'$

$$s^2 c = 24.3\, p_c' - 1.93\, |f_f| \left[\frac{\dot{c}}{|\dot{c}|}\right] \quad [12a]$$

Computer Equations

The following scale factors were chosen

$$\begin{aligned} F_f \text{ (volts)} &= f_f/5, \text{ lb} \\ X \text{ (volts)} &= 2500\, x, \text{ in.} \\ Y \text{ (volts)} &= 3000\, y, \text{ in.} \\ E' \text{ (volts)} &= 5\, e', \text{ volts} \\ P_c' \text{ (volts)} &= p_c'/10, \text{ psi} \\ C \text{ (volts)} &= 10{,}000\, c, \text{ in.} \\ R \text{ (volts)} &= 10{,}000\, r, \text{ in.} \\ \tau \text{ (computer time)} &= 100\, t, \text{ actual time} \\ \Psi \text{ (volts)} &= 10{,}000\, \psi, \text{ in.} \end{aligned}$$

Using these scale factors the system equations as set up on the computer are

$$E' = 0.915\,[R - C] \quad [1a]$$

$$\frac{Y}{E'} = \frac{0.45\, K_A}{23.2 \times 10^{-4} S^3 + 3.18 \times 10^{-3} S + 2} \quad [2a]$$

$$\frac{X}{Y} = \frac{1.29}{8.62 \times 10^{-4} S^3 + 3.96 \times 10^{-3} S + 1} \quad [9a]$$

$$0.1 S P_c' = \frac{0.188 X \sqrt{40 - P_c} - 0.1\, SC}{4 \times 10^{-3} \Psi} \quad X \geq 0$$

$$0.1 S P_c' = \frac{0.188 X \sqrt{P_c} - 0.1 Sc}{4 \times 10^{-3} \Psi} \quad X \leq 0 \quad [15a]$$

$$0.01\, S^2 c = 2.43\, P_c' - 0.0965\, F_f \left[\frac{SC}{|SC|}\right] \quad [12a]$$

Appendix 2

Force-Velocity Curves

Pressure versus flow curves can be plotted for a given valve arrangement in order to visualize more easily the region of valve operation in which the linearized flow equations apply. An insight into valve-cylinder operation also can be gained from force-velocity curves. The force-velocity curves are proportional to the pressure-flow curves. The proportionality factor depends on the area of the cylinder. These curves show graphically the amount of nonlinearity of the force-velocity relation.

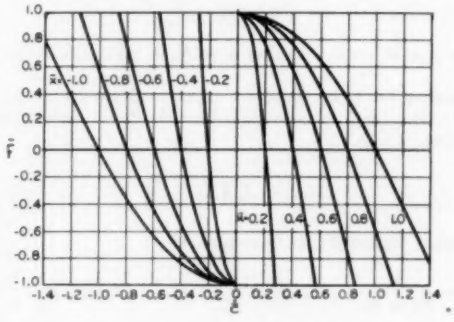


Fig. 14 Valve-Cylinder Dimensionless Force-Velocity Curves

The advantage of force-velocity diagrams is that design data can be read directly from the graph for transient or steady-state response. A force ordinate is proportional to an acceleration ordinate for a given system. The abscissa gives velocity. Thus a phase-plane plot can be directly superimposed on the force-velocity curves. The sinusoidal response is shown by an elliptical path from which can be read directly the pilot-valve openings at given velocities and total forces.

The force-velocity curves are derived in the following manner: Equation [12] can be written as

$$f = A_e (p_s - p_s/2) \quad [12b]$$

Neglecting compressibility and combining Equations [10a], [11], and [12a] yields

$$x \geq 0, \quad A_e \dot{c} = K D_v x \sqrt{\frac{p_s}{2} - \frac{f}{A_e}}$$

$$x \leq 0, \quad A_e \dot{c} = K D_v x \sqrt{\frac{p_s}{2} + \frac{f}{A_e}}$$

or

$$\bar{x} \geq 0, \quad \dot{\bar{c}} = \bar{x} \sqrt{1 - \bar{f}}$$

$$\bar{x} \leq 0, \quad \dot{\bar{c}} = \bar{x} \sqrt{1 + \bar{f}}$$

where

$$\dot{c} = \dot{c}\frac{A_o}{KD_v x_m\sqrt{p_s/2}}, \quad \bar{x} = x/x_m, \quad \bar{f} = 2f/p_s A_o$$

This relation is plotted in Fig. 14. The first and third quadrants represent useful work done by the piston. The second and fourth quadrants represent the piston being driven by the load. Since the first and third quadrants are symmetrical, only the first quadrant will be considered in the analysis. The cylinder characteristics are quite nonlinear as the slopes of the force-velocity lines vary from 0 to ∞ and the change in force due to a change in $\bar{x}$ varies from 0 to ∞. However, the cylinder characteristics can be linearized to yield order-of-magnitude results.

Actually, the valve spool has some radial clearance between it and the valve block to provide relief from pressure binding and to decrease the bearing surface. Also, the distance between the lands of the spool is slightly greater than the distance between ports. Considering these clearances a new force-velocity curve showing only the first quadrant is plotted in Fig. 9. The only effect of these clearances is a slight decrease in velocity and in force obtainable for a given valve opening.

Discussion

J. L. Shearer.[6] The emphasis on this paper is on analysis and analytical evaluations rather than on the results of detailed experimental work. It does appear that additional analysis of the system that was built and tested would be a valuable contribution to the work in this field. However, based on the analytical work alone, this paper makes a real contribution to the field of hydraulic servomechanisms because the analysis does reveal a good quantitative picture of the many problems involved in the design and development of fast reliable hydraulic servomechanisms.

It is interesting to note that the major lag in this system occurs in the torque motor and associated electronic amplifier and that this is the most severe limitation to attaining a really fast response. It would appear that if a faster response is required for a system of this kind, then a major effort would be required in improving further the dynamic response of the electromechanical transducer (i.e., torque motor). It is especially true in machine-tool control drives, where coulomb friction and similar types of nonlinear friction occur in the sliding ways and bearings of the machine, that a ramp input is a very important type of input to use during an analog study. The reason for this is that with a ramp input the output velocity has a steady component and, if this steady component is high enough, variations that occur due to an oscillatory tendency or load changes will not cause the velocity to decrease to zero and change sign. The result is that during a ramp input the coulomb friction, which depends largely on the sign of the velocity and very little on the magnitude of the velocity, does not provide the stabilizing effect that it does when frequency-response or transient-response measurements are made.

[6] Assistant Professor of Mechanical Engineering, Massachusetts Institute of Technology, Cambridge, Mass. Mem. ASME.

Nonlinear Analog Study of a High-Pressure Pneumatic Servomechanism[1]

By J. L. SHEARER,[2] CAMBRIDGE, MASS.

A detailed analog simulation was made in order to evaluate the effects of nonlinear valve characteristics, nonlinear ram-chamber compliance, and coulomb friction in the ram on the dynamic performance of a high-pressure pneumatic servomechanism that had been studied previously with a linearized analysis. This analog study revealed that a major part of the discrepancy between measured frequency-response characteristics and the frequency-response characteristics computed from a linearized analysis is caused by coulomb friction in the ram. Although the system is decidedly unsymmetrical when the ram moves near one end of its cylinder, the overall dynamic characteristics are nearly the same as when the ram moves near its center position. Thus, a simple linearized analysis of ram-chamber characteristics, which is possible only when the ram is near its center position, is applicable throughout most of the operating range of the ram. The nonlinear characteristics of the control valve did not seem to be significant in the problem because large pressure differences across the ram never were required to drive the mass load.

Nomenclature

The following nomenclature is used in the paper:

A = ram area, sq in.
C_c = capillary-resistance coefficient, in^5/lb-sec
C_p = specific heat at constant pressure, for air: 8.65×10^9 sq in/sec^2, deg R
C_v = specific heat at constant volume, for air: 6.18×10^9 s in/sec^2, deg R
D = derivative with respect to time, d/dt, sec^{-1}
d/dt = derivative with respect to time, sec^{-1}
e_1 = first input signal, volts
e_2 = second input signal, volts
e_3 = output signal, volts
e_g = input (grid) voltage to high-gain amplifier, volts
F_c = coulomb-friction force, lb
F_i = limiting (maximum) value of F_c, lb
F_r = ram pressure force, lb
f_v = functional relationship between valve flow W, valve position X, and load pressure P
g = acceleration due to gravity, 386 in/sec^2
k = ratio of specific heats, C_p/C_v, for air: 1.4
k_a = amplifier gain, volts/volt
k_f = feedback taper, in/in.
k_i = input taper, in/in.
k_s = over-all gain of electromechanical servomechanism and potentiometer, dR_1/de_2, ohms/volt
L = external load force, lb
m = load mass, lb-sec^2/in.
P = absolute pressure, psi
R = gas constant, for air: 2.47×10^5 sq in/sec^2, deg R
R_1 = potentiometer resistance, ohms
R_2 = feedback resistance, ohms
T = temperature, deg R
t = time, sec
V = volume, cu in.
W = weight rate of flow, lb/sec
X = valve position, in., zero when $W_a = W_b = 0$ and $P_a = P_b$
Y = ram position, in., zero when $V_a = V_b$
Z = position of input taper, in.
ϕ = phase shift between input and output, radians
ω = frequency, radian/sec

Frequently used subscripts:

a = a end of system
b = b end of system
e = exhaust
i = initial valve (except in k_i)
s = supply
t = stabilizing tank

Introduction

Hydraulic fluids have been used widely as the working medium in systems developed to control the motion of mass loads. Recent development of pneumatic control systems[3,4] has demonstrated that compressed gases can be used effectively for the continuous control of motion of mass loads.

The schematic diagram, Fig. 1, shows a positioning servomechanism employing a valve-controlled pneumatic servomotor to drive a mass load and external load force with a simple feedback mechanism to sense output position Y and to stroke the valve in such a manner that the pneumatic servomotor will attempt to provide changes in Y that are proportional to a low-energy-level input motion Z. Earlier papers[3] and a thesis investigation[5] by the author discuss the pneumatic processes involved in the operation of a system like that of Fig. 1, and a linearized analysis was employed to demonstrate how stabilizing resistances and stabilizing tanks could be used effectively to provide damping in the servomotor.

When the complete servomechanism was studied experimentally in the laboratory by means of frequency-response measurements, the output motion was not sinusoidal at all frequencies and the measured amplitude and phase characteristics were somewhat different from those computed from a linearized analysis. In

[1] The work reported in this paper was carried out at the Massachusetts Institute of Technology as part of an ScD thesis investigation entitled "Continuous Control of Motion With Compressed Air."

[2] Assistant Professor of Mechanical Engineering, Massachusetts Institute of Technology. Mem. ASME.

Contributed by the Instruments and Regulators Division of The American Society of Mechanical Engineers and presented at the ASME-AIEE Conference on Nonlinear Control Systems, Princeton, N. J., March 26–28, 1956.

Note: Statements and opinions advanced in papers are to be understood as individual expressions of their authors and not those of the Society. Manuscript received at ASME Headquarters, December 28, 1955. Paper No. 56—IRD-1.

[3] "Study of Pneumatic Processes in the Continuous Control of Motion With Compressed Air—Parts I and II," by J. L. Shearer, Trans. ASME, vol. 78, 1956, pp. 233–249.

[4] "Tie Simplicity to Power With Pneumatic Servomechanisms," by H. Levenstein, *Control Engineering*, vol. 2, June, 1955, pp. 65–70.

[5] "Continuous Control of Motion With Compressed Air," by J. L. Shearer, ScD thesis, Department of Mechanical Engineering, Massachusetts Institute of Technology, Cambridge, Mass., 1954.

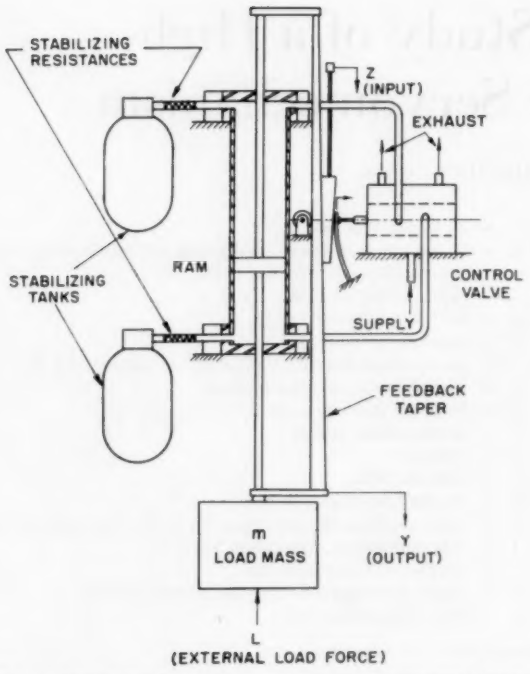


FIG. 1 SCHEMATIC DIAGRAM OF PNEUMATIC POSITIONING SERVOMECHANISM

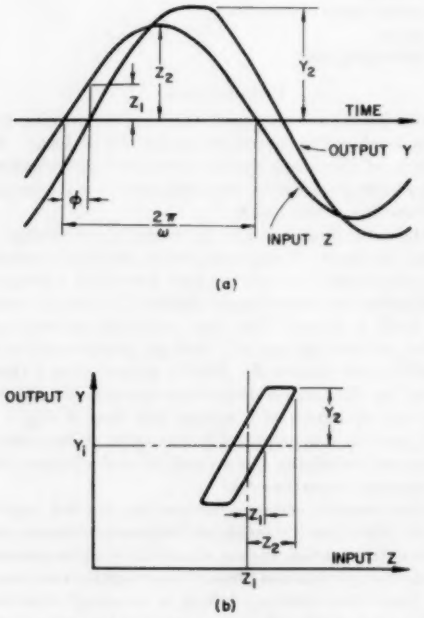


FIG. 2 SKETCHES OF INPUT AND OUTPUT WAVE FORMS
a) Input Z and output Y as functions of time. (b) Output Y as functions of sinusoidally varying input Z.]

particular, the ram came to rest when its direction of motion changed in the way shown qualitatively in Fig. 2(*a*) when the input was varied sinusoidally. This dwell in output motion, which was observed with a fairly wide range of input frequencies, appears more distinctly when the output is plotted against the input sinusoid, as shown in Fig. 2(*b*). A small amount of coulomb friction was found from static measurements of the ram, and the dwell in output motion tentatively was attributed to this coulomb friction. It was apparent, however, that either an exhaustive experimental study or a detailed analog study would be required to show quantitatively how the various known nonlinearities in the system contributed to system performance. Estimates indicated that a thorough experimental study, in which many important system parameters were changed, would be much more costly than an analog study. Although a digital computer was employed in a check solution intended to insure the validity of the analog results, the long computing time and limited information-storage facilities then available made a thorough digital-computer study inadvisable. All of the required analog-computer components were available in the Generalized Computer of the Dynamic Analysis and Control Laboratory (D.A.C.L.) of the Massachusetts Institute of Technology.

BASIC EQUATIONS AND FUNCTIONAL RELATIONSHIPS

Valve. The characteristics of the control valve are discussed in detail,[3,5] and Fig. 3 is a graphical representation of the relationship between the weight rate of flow toward one end of the ram W_a, valve position X, and ram-chamber pressure P_a. The flow toward

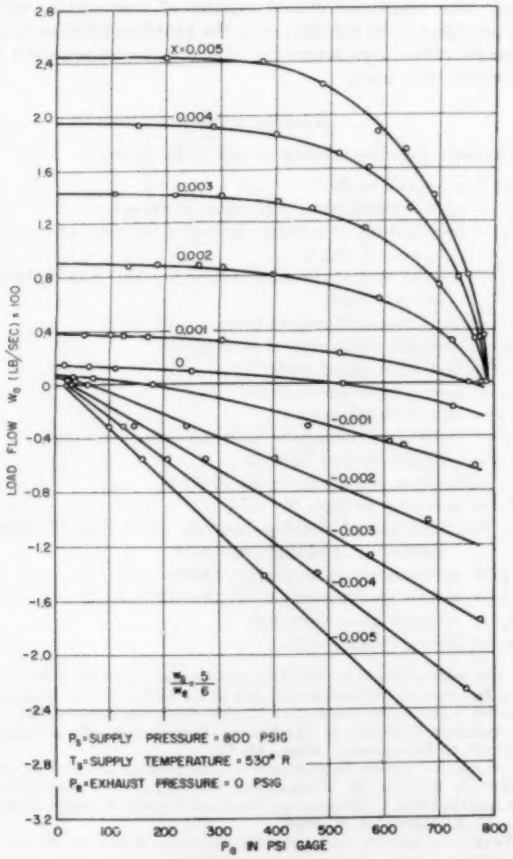


FIG. 3 MEASURED PRESSURE-FLOW CHARACTERISTICS OF NOMINAL CLOSED-CENTER SLIDING-PLATE VALVE

the other end of the ram W_b is related to valve position X and the other ram-chamber pressure P_b by an identical set of curves with only the sign of the valve position changed. In other words

$$W_a = f_v(X, P_a) \quad [1]$$

$$W_b = f_v(-X, P_b) \quad [2]$$

where f_v denotes the functional relationship given by the curves.

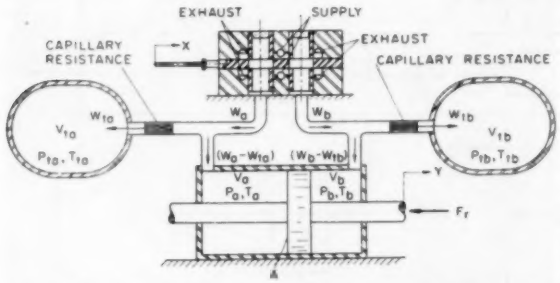


FIG. 4 SCHEMATIC DIAGRAM SHOWING IMPORTANT VARIABLES IN PNEUMATIC SYSTEM

Ram Chambers. Applying the energy equation to the ram chambers, as shown in Fig. 4, gives

$$(W_a - W_{ta})T_s - \frac{gP_a}{C_p}\frac{dV_a}{dt} = \frac{g}{kR}\frac{d}{dt}(P_aV_a) \quad [3]$$

$$(W_b - W_{tb})T_s - \frac{gP_a}{C_p}\frac{dV_b}{dt} = \frac{g}{kR}\frac{d}{dt}(P_bV_b) \quad [4]$$

where

W_{ta} = weight rate of flow into tank at a end, lb/sec
W_{tb} = weight rate of flow into tank at b end, lb/sec
T_s = temperature of gas supply, deg R
g = acceleration due to gravity, 386 in/sec²
C_p = specific heat of gas at constant pressure, sq in/sec²—deg R
V_a = volume of chamber in a end of ram, cu in.
V_b = volume of chamber in b end of ram, cu in.
t = time, sec
k = ratio of specific heat at constant pressure to specific heat at constant volume, for air: 1.4
R = gas constant, for air: 2.47×10^5 in²/sec², deg R

Stabilizing Resistances. The flow through the stabilizing resistances is obtained by using the expression for flow of a gas through a capillary resistance

$$W_{ta} = \frac{gC_c}{2RT_s}(P_a^2 - P_{ta}^2) \quad [5]$$

$$W_{tb} = \frac{gC_c}{2RT_s}(P_b^2 - P_{tb}^2) \quad [6]$$

where

C_c = capillary-resistance coefficient, in⁵/lb-sec
P_{ta} = pressure in tank at a end, psi
P_{tb} = pressure in tank at b end, psi

Reasonably good approximations for Equations [5] and [6] are given by

$$W_{ta} = \frac{gC_cP_a}{RT_s}(P_a - P_{ta}) \quad [7]$$

$$W_{tb} = \frac{gC_cP_b}{RT_s}(P_b - P_{tb}) \quad [8]$$

Stabilizing Tanks. Applying the energy equation to the stabilizing tanks gives

$$W_{ta}T_s = \frac{gV_{ta}}{kR}\frac{d}{dt}(P_{ta}) \quad [9]$$

$$W_{tb}T_s = \frac{gV_{tb}}{kR}\frac{d}{dt}(P_{tb}) \quad [10]$$

Ram and Mass Load. Application of Newton's second law to the ram and mass load yields

$$(P_a - P_b)A = m\frac{d^2Y}{dt^2} + L + F_c \quad [11]$$

where

A = ram area, sq in.
m = load mass, lb sec²/in.
Y = ram position, in., zero when ram is in center of cylinder
L = external load force, lb
F_c = coulomb-friction force, lb

Ram Volumes. The ram volumes V_a and V_b are related to Y by

$$V_a = V_i + AY \quad [12]$$

$$V_b = V_i - AY \quad [13]$$

where Y is measured from the point where $V_a = V_b = V_i$.

Feedback. The valve motion X is related to the input motion Z and the output motion Y by

$$X = k_iZ - k_fY \quad [14]$$

where

k_i = input taper, in/in.
Z = position of input taper, in.
k_f = feedback taper, in/in.

Coulomb Friction. The coulomb-friction force is constant (independent of ram velocity) when the ram is moving, and when the ram is motionless, this friction force is equal to the sum of all other forces acting on the ram

$$F_c = \frac{\frac{dY}{dt}}{\left|\frac{dY}{dt}\right|}F_l \quad \left[\frac{dY}{dt} > 0\right] \quad [15a]$$

$$F_c = (P_a - P_b)A - L \quad \left[\frac{dY}{dt} = 0\right] \quad [15b]$$

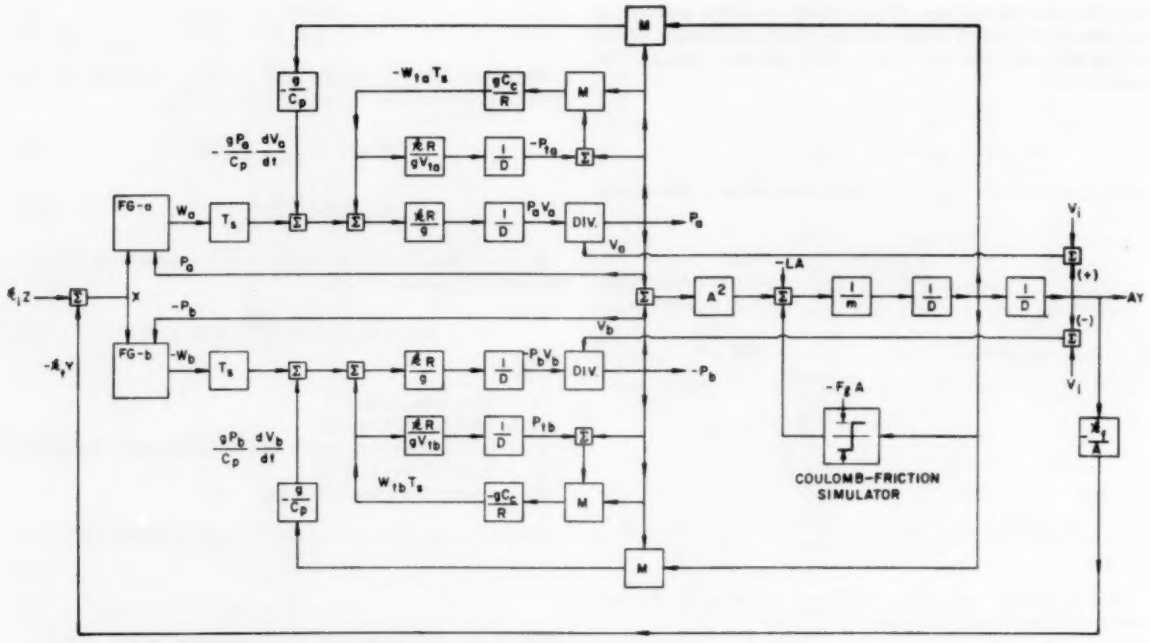
where

F_l = limiting (maximum) value of F_c, lb

DESCRIPTION OF ANALOG SYSTEM

Block Diagram of Complete System. The block diagram shown in Fig. 5 represents the complete set of equations describing the system and gives an over-all picture of the operation of the system. It also reveals the role played by each physical characteristic of the system and the many interactions within the system.

Two of the D.A.C.L. function generators were employed to simulate the valve characteristics. Fig. 6 is a view of one three-dimensional profile and reading head that was used to generate the valve flow as a function of valve position and ram pressure. Position-type electromechanical servomechanisms provide means



(M denotes multiplier; DIV denotes divider; FG denotes function generator; $1/D$ denotes integration with respect to time; Σ denotes summation.)

FIG. 5 BLOCK DIAGRAM OF COMPLETE PNEUMATIC SERVOMECHANISM

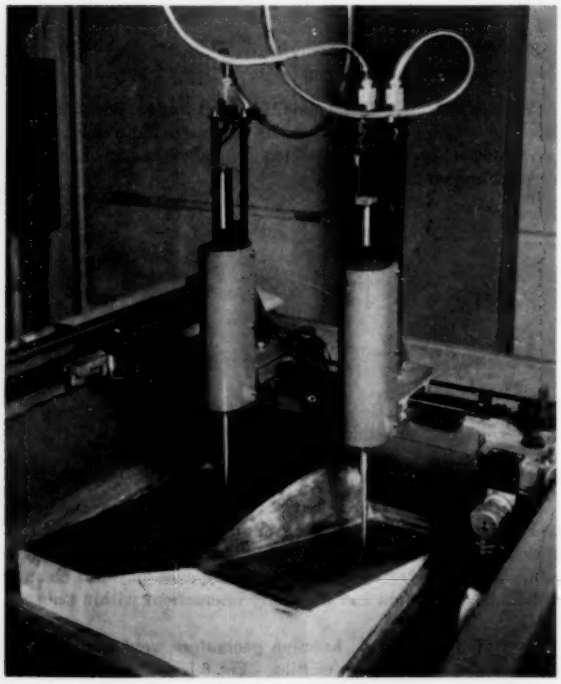


FIG. 6 VIEW OF FUNCTION GENERATOR SETUP TO SIMULATE a END OF CONTROL VALVE AND TO PERFORM MULTIPLICATIONS

[$P_a \times dV_a/dt$ and $P_a \times (P_a - P_{ta})$.]

of introducing a valve-displacement signal X and a ram-pressure signal P_a by positioning the reading head along the cross carriage and by positioning the cross carriage along the machine frame, respectively. A linear potentiometer in the reading head delivers a signal proportional to the height of the three-dimensional profile which represents flow toward the ram. Also shown is a constant-slope profile and accompanying reading head clamped to the cross carriage. This reading head contains two potentiometers: One is excited with a voltage representing dV_a/dt, and the other is excited with a voltage representing $(P_a - P_{ta})$. Thus the voltages picked up by the two potentiometer wipers represent P_a times dV_a/dt and P_a times $(P_a - P_{ta})$, respectively. A similar arrangement simulates the situation at the other end of the ram and valve.

Each of the quotients P_aV_a/V_a and $-P_bP_b/V_b$ is obtained with an electromechanical dividing system like that shown in Fig. 7.

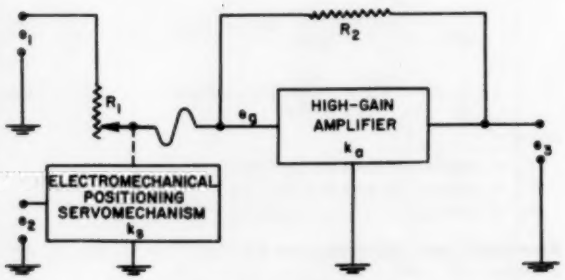


FIG. 7 SCHEMATIC DIAGRAM OF ELECTROMECHANICAL SYSTEM USED TO DIVIDE e_1 BY e_2

[$e_3 = -(R_1/k_s)(e_1/e_2)$ when $(R_2 + R_1)/k_a = R_1$.]

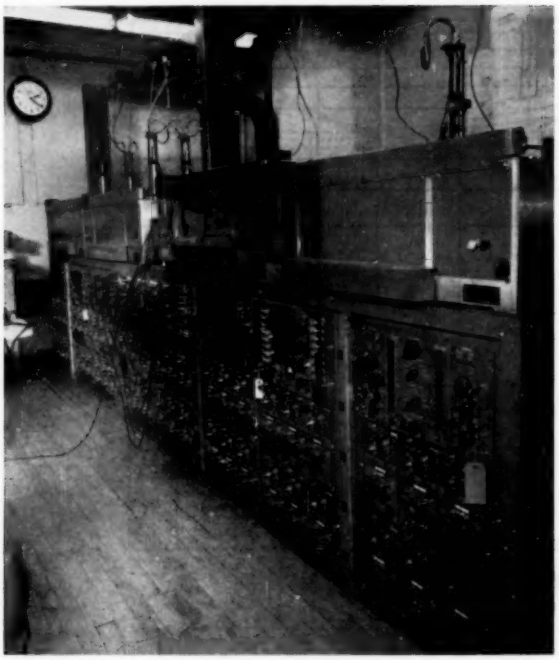


FIG. 8 OVER-ALL VIEW OF ANALOG-COMPUTER EQUIPMENT

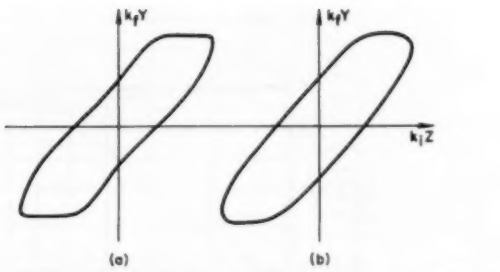


FIG. 9 COMPARISON OF FREQUENCY-RESPONSE TEST OF EXPERIMENTAL SYSTEM WITH SIMULATED TEST ON ANALOG
[(a) Y versus Z of analog at 1.0 cps. (b) Y versus Z of experimental system at 1.0 cps.]

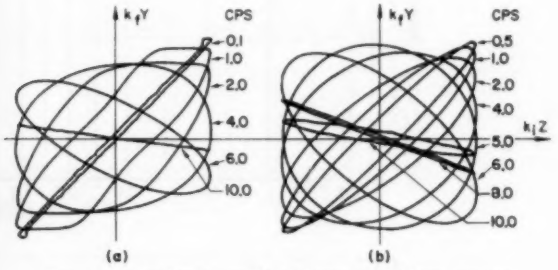


FIG. 10 TWO SERIES OF TESTS SIMULATED WITH ANALOG COMPUTER
[(a) With coulomb friction. (b) Without coulomb friction.]

Conventional summing, coefficient, and integrating circuits are used with chopper-stabilized d-c amplifiers to achieve these necessary operations. Coulomb friction was simulated with a high-gain amplifier and precision limiter. Fig. 8 is a view of the complete analog setup. Each system variable is represented by an electric signal having a voltage proportional to the value of the variable. The time scale used in the greater part of this study was 8 to 1; in other words, the integrators were 8 times slower than real-time integration.

RESULTS OF DYNAMIC SYSTEM STUDIES

The system studied in greatest detail with the analog was the one which had been measured in the laboratory. The control-valve characteristics are shown in Fig. 3; supply pressure P_s was 800 psig; supply temperature T_s was 530 R; half ram volume V_t was 30 cu in.; ram area A was 4.34 in.^{2}; stabilizing tank volume V_t was 130 cu in.; stabilizing-resistance coefficient C_s was 4.46 in^5/lb-sec; maximum coulomb friction F was 13.0 lb; load mass m weighed 70 lb; feedback taper k_f was 0.022 in/in.; and input taper k_i was 0.25 in/in. The result of an analog simulation is compared in Fig. 9 to the measured response of the experimental system when the input Z was varying sinusoidally with an amplitude of ± 0.050 in.

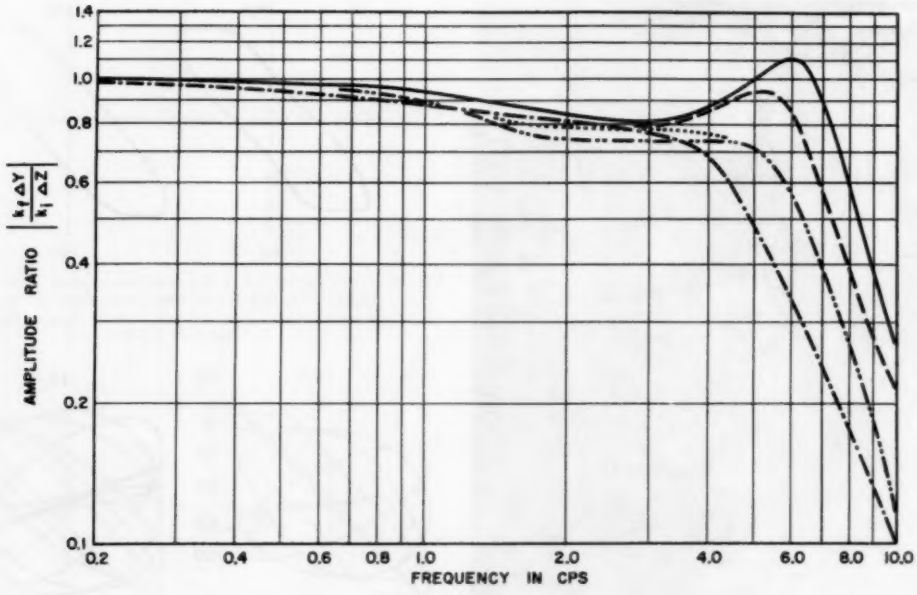
Several series of steady-state frequency-response tests were simulated on the analog in order to determine differences between the analog system and the real system, and to observe the effects of varying the input amplitude, coulomb friction in the ram, and the operating position of the ram. Fig. 10 illustrates two of the series of tests that were simulated on the analog with the ram operating near its center position. The series of Fig. 10 (a) was made with the measured coulomb friction set into the analog, and the series of Fig. 10(b) was made with zero coulomb friction set into the analog. This shows clearly that the dwell in output motion, which was discussed earlier, is caused by coulomb friction. Each series required only a few minutes to run off. Other tests that were simulated on the analog included frequency-response measurements with the coulomb friction doubled and with the input amplitude reduced from ± 0.050 in. to ± 0.025 in. The results of all of these tests appear in Fig. 11, where the response computed from a linearized analysis[3,5] is given also for purposes of comparison. The results from tests of the experimental system are shown in Fig. 12. The greatest discrepancy between the analog simulation and the experimental system itself appears in the frequency range of 3 to 5 cps.

Series of frequency-response tests that were simulated for the ram operating near positions 2 in., 4 in., and 6 in. from its center position (maximum stroke = ± 6.5 in.) did not vary perceptibly from those made near the center position. It was observed, however, that during operation near one end of the ram cylinder, the major pressure variation occurred in the smallest ram chamber and attached stabilizing tank, an effect demonstrating the need for two tanks and resistances. The magnitude of the pressure variations was always less than 100 psi, with the result that the control valve operated in a nearly linear fashion throughout the tests. Reasonably good agreement was obtained between pressure variations in the real system and pressure variations in the analog.

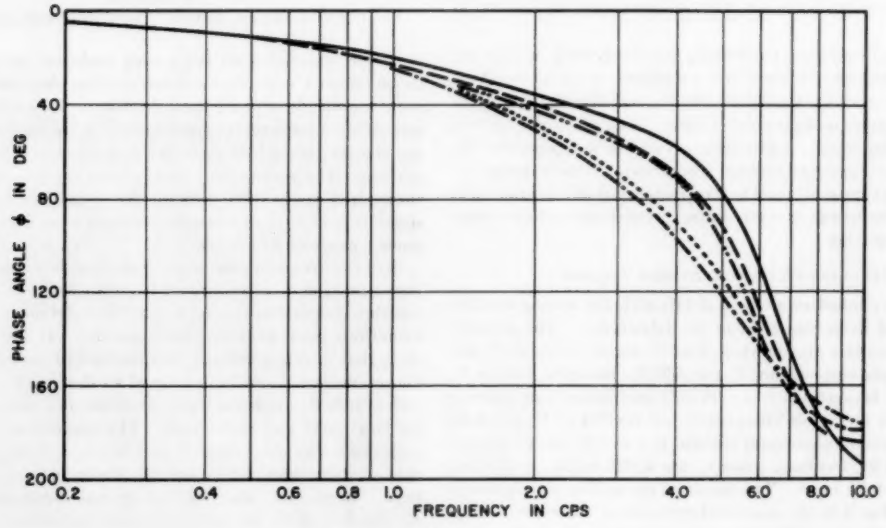
CONCLUSIONS

Although the linearized analysis does not predict actual system performance accurately, it does provide an excellent means of gaining a qualitative insight into the dynamic characteristics of the system, especially in regard to the effectiveness of the stabilizing tanks and resistances. A fact predicted by the linearized analysis is that this system is very unstable when the stabilizers are shut off. Both the real system and the analog demonstrated this fact very dramatically.

The analog proved to be a highly effective tool in attaining optimum designs for a system of this kind having different sets of requirements. A number of such designs were completed in the



(a)



(b)

——————————Computed from linearized analysis
———— ————Simulator, no friction, ±0.050-in. input
— — — — — — —Simulator, twice-measured coulomb friction, ±0.050-in. input
—— — — —— — —Simulator, measured coulomb friction, ±0.050-in. input
—— — ——— —Simulator, measured coulomb friction, ±0.025-in. input

FIG. 11 RESULTS OF SEVERAL SERIES OF STEADY-STATE FREQUENCY-RESPONSE TESTS SIMULATED WITH ANALOG COMPUTER COMPARED WITH RESPONSE COMPUTED FROM LINEARIZED ANALYSIS
[(a) Amplitude versus frequency. (b) Phase angle versus frequency.]

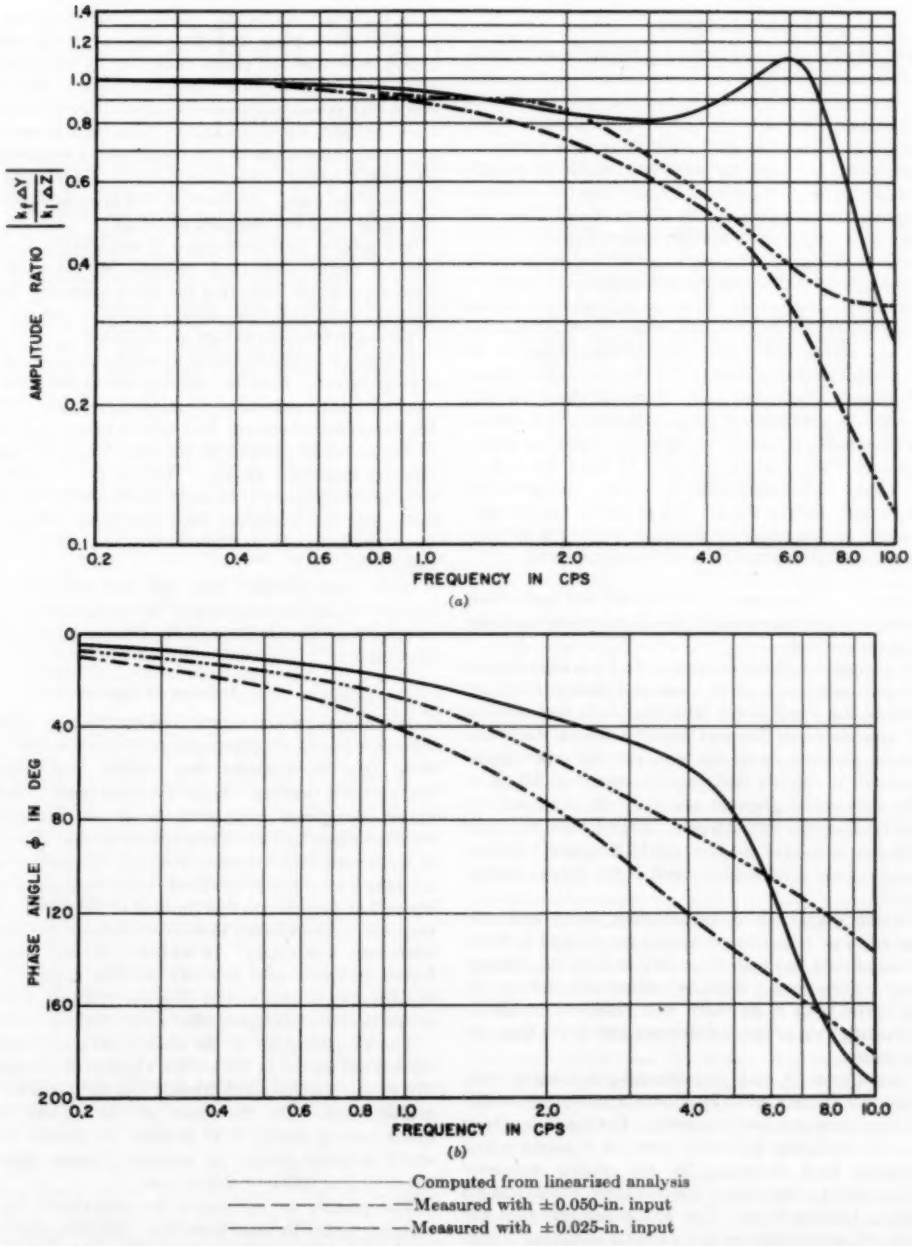


FIG. 12 RESULTS OF TWO SERIES OF STEADY-STATE FREQUENCY-RESPONSE TESTS ON EXPERIMENTAL SYSTEM COMPARED WITH RESPONSE COMPUTED FROM LINEARIZED ANALYSIS
[(a) Amplitude versus frequency. (b) Phase angle versus frequency.]

space of a few hours after the analog was set up. The cost in time, money, and effort of trying various modifications of the real system would have been many times greater.

ACKNOWLEDGMENTS

The author is indebted to many members of the Dynamic Analysis and Control Laboratory for the assistance and the counsel given in carrying out the work reported here.

Outstanding among them are: H. Mori (analog setup and operation), C. W. Gould (experimental work), J. M. Aitken (drafting), and Prof. J. A. Hrones, Director of the Laboratory.

The work described here was supported in part by the Bureau of Ordnance, U. S. Navy Department, under Contract No. NOrd 11799, with the Division of Industrial Cooperation, Massachusetts Institute of Technology.

Discussion

C. Muzzey.[6] The author's work on gas-driven servo systems is extremely interesting and he is to be commended for producing a unit which can control a sizable mass load at a very respectable frequency response. The paper presents a good picture of the analog computer setup used to study and design the system.

It would seem that the computer setup used would be capable of handling somewhat larger input signals than those reported, and that some transient responses could be obtained and compared with tests on the actual servo. On the other hand, restriction to small inputs probably would permit further simplification of the analog, particularly with respect to valve characteristics.

The penalty in equipment size in using a pneumatic rather than hydraulic working fluid is considerable. For example, if the author's servo ram has a half-stroke displacement of 30 cu in. and a maximum half stroke of 6.5 in., the piston area is 4.61 sq in. to control a 70-lb mass load. Two stabilizing tanks of considerable size also are required to make the gas servo behave. Recently we designed a hydraulic servo for a somewhat higher frequency response with a piston area of 0.75 sq in. for a 33-lb mass load. Thus, when conditions of ambient temperatures or fluid power supply dictate the use of a gas servo, the designer must be prepared to find space for a device which will be large when judged by familiar standards of hydraulic equipment.

D. V. Stallard.[7] This paper presents a lucid and penetrating study of a nonlinear servomechanism that is inherently complex and difficult to understand.

By way of comparison, friction appears to be a more serious problem in a pneumatic ram servo than in a similar hydraulic servo, because of the much lower lubricity of air and the fact that it takes appreciable air flow and time to build up the breakaway differential pressure across the ram. In the experimental pneumatic system, it appears that the ram area was 4.6 sq in. Therefore the differential pressure necessary to overcome the 13-lb coulomb friction was only 2.8 psi. If the system fluid had been oil at 800 psig instead of air, then the 13-lb coulomb friction would not have caused a noticeable dwell in the output motion during reversal.

Because friction lowers the static accuracy, many hydraulic servos of the valve or transmission type have utilized a dither signal of low amplitude and high frequency to keep the friction-loaded member broken loose. Perhaps a pneumatic servo would have a lower static error if its valve were dithered enough to keep the alternating ram pressure difference just lower than the breakaway pressure.

In some control systems with proportional-plus-integral compensation, coulomb friction actually causes a nonlinear oscillation at a low amplitude and low frequency. For example, Haas[8] has described an oscillating hydraulic servo on a machine tool with considerable load friction. In his analog computer study, Haas simulated a breakaway friction force which was 1.3 times the sliding friction force. The writer suspects that the author's pneumatic servo might exhibit a similar nonlinear oscillation at a low frequency if proportional-plus-integral compensation were added.

In making a sinusoidal frequency-response test, special care must be taken when the output wave form is nonsinusoidal or noisy; otherwise the phase data may be very inaccurate. One good method[9] is to form a null Lissajous pattern on an oscilloscope with the output wave form and a sine wave which is shifted from the input wave by a known, adjustable phase angle. In this way, the phase angle of the fundamental frequency in the nonsinusoidal output wave form may be found with comparative accuracy and ease. By contrast, it is often misleading to measure the phase angle between zero crossings.

The analog instrumentation is remarkably thorough. However, it might have been desirable to eliminate the function generators which simulated the valve characteristics. Since the pressure variations were always less than 100 psi, the operation of the control valve was very nearly linear.

It appears preferable to compare the data of the experimental system directly with the corresponding simulator data rather than with the results of the linearized analysis. When replotted, the experimental-system data agreed reasonably well with those of the simulator, except in the high-frequency region, which is not very important anyway. Evidently the effect of increasing the input amplitude is to make the experimental system behave more like the linearized approximation, perhaps because the force necessary to accelerate the ram increases with amplitude.

It is quite probable that this and previous papers by the author will advance the state of the pneumatic art, which appears to have been retarded by a lack of understanding of the physical processes.

Author's Closure

Mr. Muzzey's comments regarding larger input signals and transient-response measurements are well taken. Much of the work that he suggests was actually performed as part of the author's doctoral thesis investigation[4]. However, careful transient-response measurements of the experimental system were not made because of limitations of time, effort, and facilities. It was found that frequency-response measurements with larger-amplitude inputs were rendered rather meaningless at frequencies above 3 or 4 cps due to flow limiting in the valve. The transient-response measurements that were made on the nonlinear analog were very interesting. In addition to step inputs, ramp-type inputs also were used to study stability during conditions when the direction of ram motion does not change; that is, when there are no beneficial damping effects from coulomb friction.

The simplification of the analog for small input signals was carried out earlier in the author's papers on pneumatic processes (3), except that coulomb friction was not included in this earlier analog work. On the basis of the results from the nonlinear analog study, it is possible to predict that an analog which is linear except for coulomb friction should provide an effective simulation of this system.

The penalty in equipment size mentioned by Mr. Muzzey certainly exists in some instances. In other instances, however, the penalty in size is negligible, depending on the size of the mass load to be driven and depending on speed-of-response requirements.

The comments offered by Mr. Stallard are pertinent and well made and no additional rebuttal seems necessary.

---

[6] Flight Research Department, Cornell Aeronautical Laboratory, Inc., Cornell University, Buffalo, N. Y.

[7] Research Engineer, Servomechanisms Laboratory, Massachusetts Institute of Technology, Cambridge, Mass.

[8] "Coulomb Friction in Feedback Control Systems," by V. B. Haas, Jr., AIEE Transactions Paper No. 53-108, *Applications and Industry*, vol. 72, May, 1953, pp. 119-126.

[9] "Principles of Servomechanisms," by G. S. Brown and D. P. Campbell, John Wiley & Sons, Inc., New York, N. Y., 1948, pp. 317-320.

A Dual-Mode Damper-Stabilized Servo

By J. JURSIK,[1] J. F. KAISER,[2] AND J. E. WARD[3]

The highly oscillatory response of an inertia-damper stabilized servo to large step inputs can be improved greatly by changing, for large error, the coupling torque between the damper and the drive motor. This paper presents the results obtained from an experimental model and an analog computer study of the dual-mode damper-stabilized servo. For large step inputs, this dual-mode method of operation resulted in a reduction of settling time by a factor of 5.8 and reduction of peak overshoot by a factor of 200 as compared with single-mode operation.

Introduction

INSTRUMENT servos employing inertia-damper stabilization[4] have been used extensively in recent years where high performance is required. This type of servo stabilization has several advantages; i.e., very high velocity and torque constants may be obtained along with extremely smooth and reliable operation; the servo compensation is not sensitive to changes in the a-c supply frequency; and because the amplifier can be a conventional a-c type, a system free of zero drift can be constructed. An inertia damper consists of an inertia slug coupled viscously to the servo-motor shaft, usually by means of a viscous fluid or a magnetic eddy-current coupling.

The major disadvantage of inertia-damper stabilization has been that system transient response deteriorates rapidly for inputs which cause the system error to exceed the linear range, which is quite small because of the high static system gain. This deterioration is brought about primarily by energy storage in the damper inertia during periods of nonlinear operation. Although a damper-stabilized repeater can be designed so that it always operates in the linear range during normal follow-up operation, the transient following any large abrupt change in input position which causes the repeater to slew is usually quite severe, and may limit the usefulness of the repeater in certain applications.

The purpose of this paper is to show how the large-signal performance of an instrument servomechanism employing inertia-damper stabilization can be improved by altering the coupling between the stabilizing damper and the system. This coupling can be either a simple slip clutch, or an electrically operated clutch controlled by a function of the servo error. The error-actuated clutch, although more complicated, has been found to be superior because of greater reliability. Its physical design and operation will be described in the remainder of the paper.

Description of Dual-Mode Servo

Physical Characteristics. A cross-sectional view of the dual-mode damper package attached to a servomotor is shown in Fig. 1. The damper consists of a sintered-tungsten slug mounted on bearings within a spun aluminum case filled with a silicone damping fluid. The aluminum damper case is supported by two bearings; one fixed in an extension of the motor housing and one running on the motor shaft. The clutch is composed of a small collet attached to the damper case, which can be expanded into a drum attached to the motor shaft.

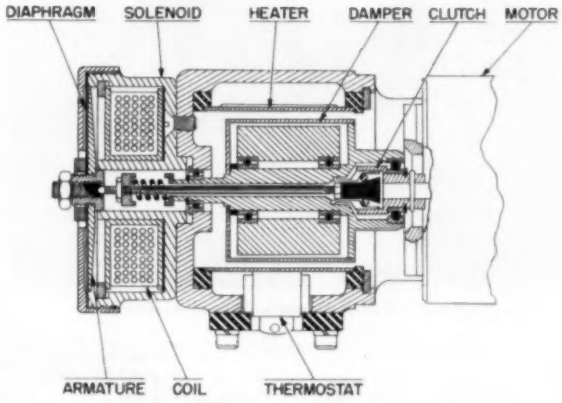


Fig. 1 Clutch-Damper Cross Section

The clutch normally is kept engaged by a spring which acts on a push rod so as to expand the collet. The means for controlling this clutch is provided by the solenoid, which releases the collet by acting on the push rod in opposition to the spring. A 4-lb force is required to release the clutch and the necessary travel is 0.005 in. The solenoid is constructed very much like a telephone receiver. It has a circular armature, which is pulled down toward the center post when the coil is energized. The return spring for the armature is provided by the thin diaphragm attached to the armature at its center. The clutch can be disengaged by approximately 2.5 watt in the solenoid coil. The operating time for the clutch is less than 10 millisec. Adjustments are provided for setting the maximum slip torque of the clutch, and the free play between the solenoid armature and the push rod. The total armature travel is established by grinding the center post during manufacture.

Because the damper is sensitive to fluid viscosity, a thermostatically controlled heater tube is included in the assembly to keep the temperature of the fluid in the damping gap at 160 F. The heater is required for airborne use, but may be dispensed with for other applications. The entire assembly is $2^3/_4$ in. long $\times$ 2 in. in diameter and has been designed to operate with a Mark 8, 400-cps servomotor. A view of the complete dual-mode package is shown in Fig. 2.

A block diagram of a typical instrument servo utilizing the dual-mode package is shown in Fig. 3. This servo will be described further at the end of the paper. As shown by the heavy arrows, the normal servo loop includes the fine synchro-control transformer, the servoamplifier, the servomotor, and the gearing to the dials and the synchro. As is usual in two-speed synchro data-transmission systems, a fine-coarse data switch is provided to switch from the fine synchro to the coarse synchro during large

[1] Engineer, Clevite Research Center, Cleveland, Ohio.

[2] Research Assistant, Servomechanisms Laboratory, Department of Electrical Engineering, Massachusetts Institute of Technology, Cambridge, Mass.

[3] Executive Officer, Servomechanisms Laboratory, Department of Electrical Engineering, Massachusetts Institute of Technology, Cambridge, Mass.

[4] "Damper-Stabilized Instrument Servomechanisms," by Albert C. Hall, AIEE Paper 49–79, December, 1948.

Contributed by the Instruments and Regulators Division of The American Society of Mechanical Engineers and presented at the ASME-AIEE conference on Nonlinear Control Systems, Princeton, N. J., March 26–28, 1956.

Note: Statements and opinions advanced in papers are to be understood as individual expressions of their authors and not those of the Society. Manuscript received at ASME Headquarters, January 11, 1956. Paper No. 56—IRD-6.

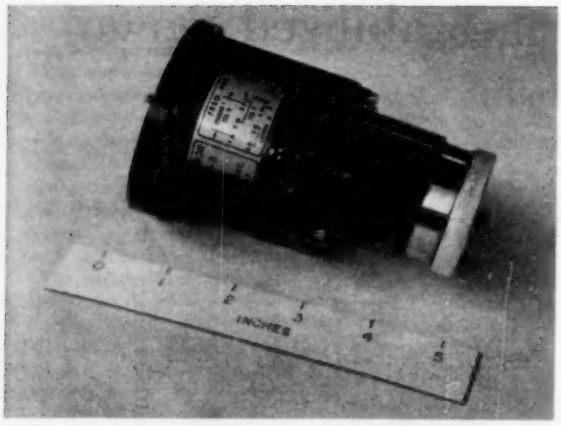


FIG. 2 MOTOR, DAMPER, CLUTCH ASSEMBLY

cuit. In the relay circuit, shown in Fig. 4, the coarse synchro voltage is amplified and applied to a full-wave detector. The d-c voltage thus derived is used to control a triode which drives a high-speed relay capable of operating in less than 1 millisec. This relay in turn supplies a direct current (approximately 100 milliamp) to operate the solenoid.

Although the relay circuit was found to operate satisfactorily in laboratory-life tests, problems in operating relays in severe environments led to the design of the magnetic-amplifier circuit, shown in Fig. 5. The coarse synchro voltage is again detected and amplified, but the relay coil in the plate circuit of the control tube is replaced by the control winding of a magnetic amplifier. The solenoid coil is operated from an a-c source by connecting it in series with a silicon diode and the load winding of the magnetic amplifier.

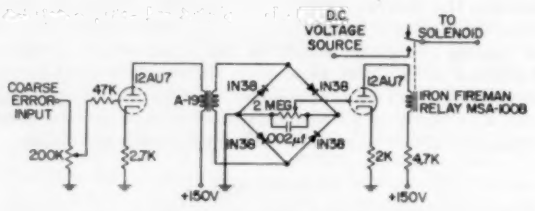


FIG. 4 RELAY CONTROL CIRCUIT

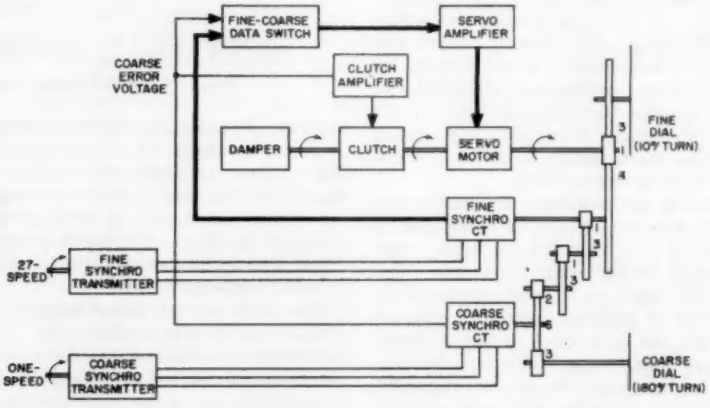


FIG. 3 BLOCK DIAGRAM OF DUAL-MODE SERVO

errors. Stabilization of the servo in the normal mode is provided by the damper connected to the servomotor, and the servoamplifier provides only voltage amplification.

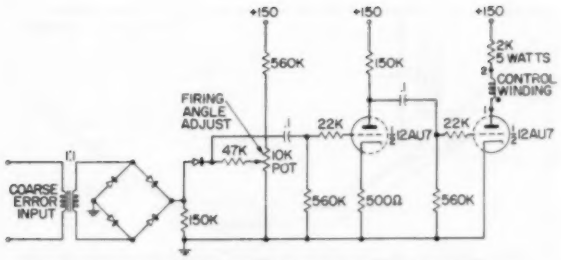
In dual-mode operation, when the coarse error exceeds a certain value (which is usually not the same as the value for fine-coarse switching), the clutch is actuated so as to disengage the damper from the servomotor. The servomotor, freed of the inertia of the damper, can then rapidly accelerate and slew the system to the new synchronization point. As soon as the error is again reduced to the switching angle, the damper is again connected to the servomotor, and acts as an inertia brake to bring the system rapidly into synchronization. Usually the overshoot is so small that the system does not again leave the linear range after once entering it. The reason that the value of coarse error for clutching may be different from the value for fine-coarse switching is that there is a separate criteria for setting each one. The clutching angle is based on dynamic considerations to be described in the next portion of the paper. The fine-coarse switching angle is determined by the gear ratio between the fine and coarse synchros.[5]

A number of control circuits to operate the solenoid from the coarse error have been designed and tested. The two most satisfactory are a relay-control circuit and a magnetic-amplifier cir-

Performance. As has been stated, the transient performance of a damper-stabilized servo for large signals, such as those encountered in synchronizing to the new position, is sometimes undesirable. A typical large-signal transient for single-mode operation of the system[6] of Fig. 3 is shown in the upper portion of Fig. 6. It will be noted that the system has an 80 per cent initial overshoot and requires at least 6 complete cycles to reach and stay within the linear range of operation.

[5] "How to Design Speed Switching Circuits," by Basil T. Barber, *Control Engineering*, November, 1954, pp. 50–53; December, 1954, pp. 33–36.

[6] The system, which was used as an experimental model for this development, is described in detail in the Appendix.



MAGNETIC AMPLIFIER CONTROL CIRCUIT FOR DUAL-MODE SERVO

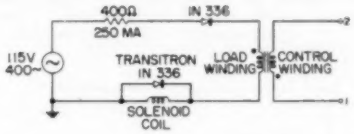


FIG. 5 MAGNETIC-AMPLIFIER CIRCUIT

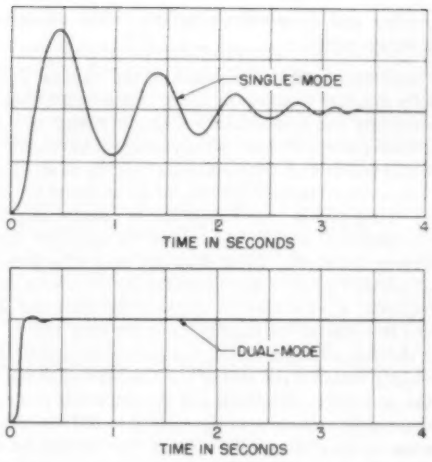


FIG. 6 EXPERIMENTAL SERVO RESPONSE TO 53-DEG STEP

With the dual-mode system in operation, the transient shown in the lower part of Fig. 6 is obtained. In this mode the system initially reaches the desired position more rapidly because of the reduced system inertia with the damper disconnected. Reconnection of the damper just prior to crossing the desired position results in an exceedingly high rate of deceleration which prevents the system from overshooting more than a fraction of 1 per cent. Once within the linear range, the servo completes synchronization rapidly because of the wide bandwidth of normal servo loop. The system completely settles in less time than is taken to reach the first overshoot in single-mode operation.

DESIGN CONSIDERATIONS

Before delineating the design criteria, it is necessary to consider the type of input which the system is to follow. The input consists of a series of ramps of variable slope separated by large step changes. Straight damper compensation gives satisfactory performance during the ramp input. The second mode of operation is added only to improve the system response to the large step changes.

The dual-mode system offers great improvement in response to large steps as has been shown in Fig. 6. A second approach consists of the use of a slip clutch to connect the motor to the damper. These three systems, single-mode, dual-mode, and slip-clutch, are compared in Fig. 7 as to their time response to various size steps. The responses correspond to three step sizes representing linear operation, moderate saturation, and hard saturation of the error amplifier. The single-mode step responses show the degeneration in response with increasing step size. The improvement due to dual-mode operation is realized for large steps only, i.e., those sufficient in size to cause damper declutching. The slip-clutch system alters all step responses that demand a damper driving torque in excess of the slip-clutch limit. This effect is seen in the step response for steps as small as 0.02 deg. A marked improvement over both the single and dual-mode systems is noted in the step response for steps in the range of a few degrees in size.

From Fig. 7 it appears that the slip-clutch system gives the best response over a wide range of input sizes. However, it was found experimentally, as shown in Fig. 8, that the response is affected greatly by the slip-clutch torque setting. For small values of slip torque the small-step response became more oscillatory as a result of reduced effective damping. For larger values of slip torque the large-step response became more oscillatory. The effectiveness of

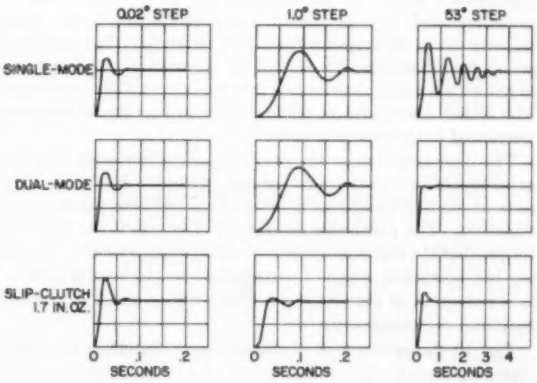


FIG. 7 RESPONSE OF EXPERIMENTAL SERVO TO VARIOUS SIZE STEPS FOR SINGLE-MODE, DUAL-MODE, AND SLIP-CLUTCH OPERATION

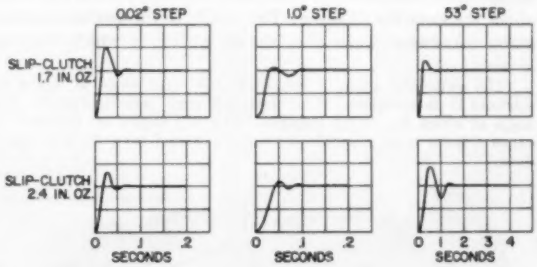


FIG. 8 EFFECT OF CHANGE IN SLIP-CLUTCH SETTING ON RESPONSE TO STEP INPUTS FOR EXPERIMENTAL SERVO

the slip-clutch coupling hence depends on the maintaining of a constant value of slip-torque. This then becomes a stringent requirement on the mechanical design of the slip clutch.

The main factors affecting the design of the dual-mode system are:

1 Solenoid pull-in and drop-out times.

2 Type of signal; i.e., what function of the error is used to control clutch actuation.

3 Driving and drag torques of the clutch connecting the damper to the motor.

The objectives behind disconnecting the damper during the large-error state are to allow the motor to accelerate most rapidly to slew velocity and to minimize the kinetic energy stored in the damper during slew. To take full advantage of the clutched mode the solenoid pull-in and drop-out times must be as small as possible. (The action of the solenoid dropping out causes the clutch to engage.) Long pull-in time results in the damper acquiring appreciable rotational energy which must be absorbed during the final settling transient. Long drop-out time effectively means that the solenoid must be deactivated while the error is still large, thus requiring a wide zone of single-mode operation if clutch actuation depends on the magnitude of the error alone.

With the damper disconnected, the system can accelerate to full slew velocity within 3 per cent of the total travel of the system. Since the probability distribution of the amplitude of step inputs is approximately uniform over the range 0 to 100 per cent of total system travel, the system will reach full slew velocity for about 94 per cent of the step inputs. Therefore the magnitude of error can be used as the sole means to actuate the solenoid. The clutch reengagement time can be compensated for by increasing the switching angle[7] by an amount equal to the maximum output velocity times the drop-out time of the solenoid. The addition of error-rate information to control the solenoid would be effective in reducing settling time for less than 5 per cent of the step inputs. The added complexity of adding an error-rate signal seems unwarranted in the system studied.

The determination of the optimum clutching angle was done experimentally with the results shown in Fig. 9. Minimum amplitude of the first overshoot is used as the criterion for this determination. The parameter for the set of curves is the maximum torque that the clutch can transmit in the engaged state. Increasing the maximum value of transmitted torque results primarily in a reduction of overshoot amplitude with a small reduction in optimum clutching angle.

Fig. 10 shows the final transients after clutching for various clutching angles. The clutch torque slip level of 5.8 in. oz is 2.4 times the maximum stall torque of the drive motor. It is interesting to note that the change in clutching angle affects only the amplitude of overshoot, with the final transient settling time being about the same for all cases. The results of the analog-computer study corroborated the data shown in Fig. 9, which indicates

[7] The switching angle is defined as the error angle at which the solenoid is de-energized. The clutching angle is defined as the error angle at which the clutch engages. The two angles are different because a finite time is required for the solenoid force to build up or decay.

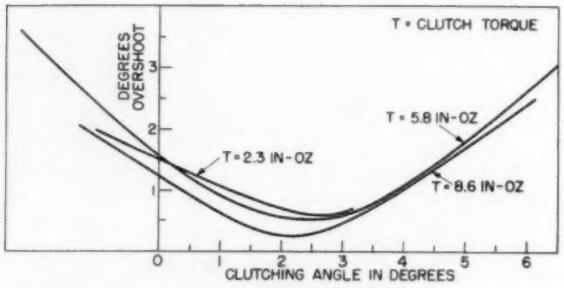


FIG. 9 EFFECT OF VARIATION IN CLUTCH TORQUE ON OPTIMUM CLUTCHING ANGLE AND OVERSHOOT FOR A 53-DEG STEP WITH EXPERIMENTAL DUAL-MODE SERVO

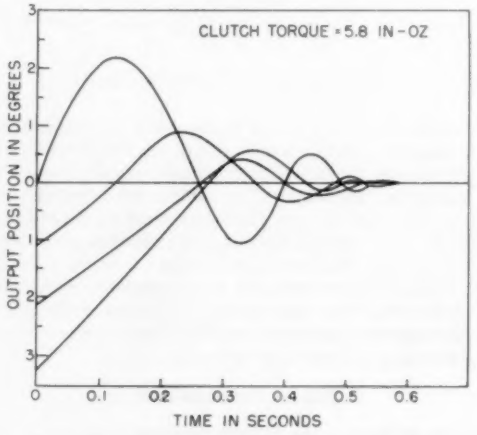


FIG. 10 FINAL TRANSIENTS AFTER CLUTCHING FOR EXPERIMENTAL DUAL-MODE SERVO FOR VARIOUS CLUTCHING ANGLES WITH 53-DEG STEP INPUT

that there is a well-defined minimum. Also studied on the analog computer was the effect of motor velocity at the time of clutching on the optimum clutching angle. These results, which are pertinent only for those cases where the motor has not reached slew velocity, are shown in Fig. 11. The computer results also indicate that settling time is approximately a minimum for the clutching angle that gives minimum overshoot.

For a fixed clutch slip torque of 5.8 in. oz, the optimum clutching angle was determined (see Fig. 10) to be 2.1 deg. To this

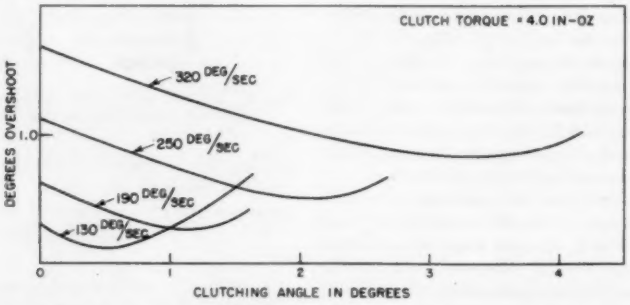


FIG. 11 EFFECT OF VARIATION IN MOTOR VELOCITY AT CLUTCHING UPON OPTIMUM CLUTCHING ANGLE AND OVERSHOOT MAGNITUDE

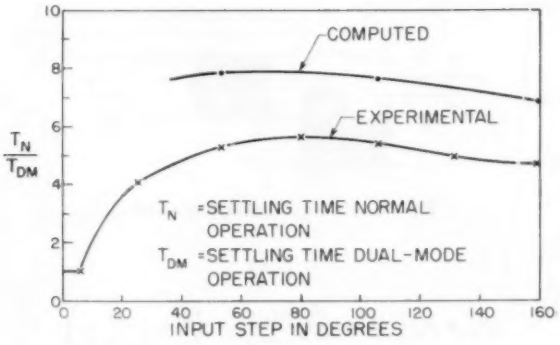


FIG. 12 SETTLING TIME REDUCTION FOR DUAL-MODE OPERATION

APPLICATIONS

Good response to both large and small step inputs, along with the fast settling times provided by dual-mode operation make the solenoid-actuated clutch attractive in spite of its added complexity to the system. It is ideally suited for improving the synchronizing performance in a damper-stabilized synchro data repeater.[8] Fig. 14 shows a data repeater with the dual-mode package as shown in Figs. 1 and 2 installed.

This instrument has a gain-crossover frequency of 40 cycles per sec (cps), a velocity constant[9] in excess of 10,000 sec^{-1}, a static accuracy of 0.02 deg, and a maximum slew velocity of 320 deg per sec. The repeater can follow smoothly at speeds as low as 0.01 deg per sec., giving a total speed range greater than 30,000 to 1. By adding the dual-mode feature, the settling time for a 180-deg step input is reduced from 5 to 1.1 sec, hence keeping total synchronizing time under 1.1 sec for any size step. The clutch mecha-

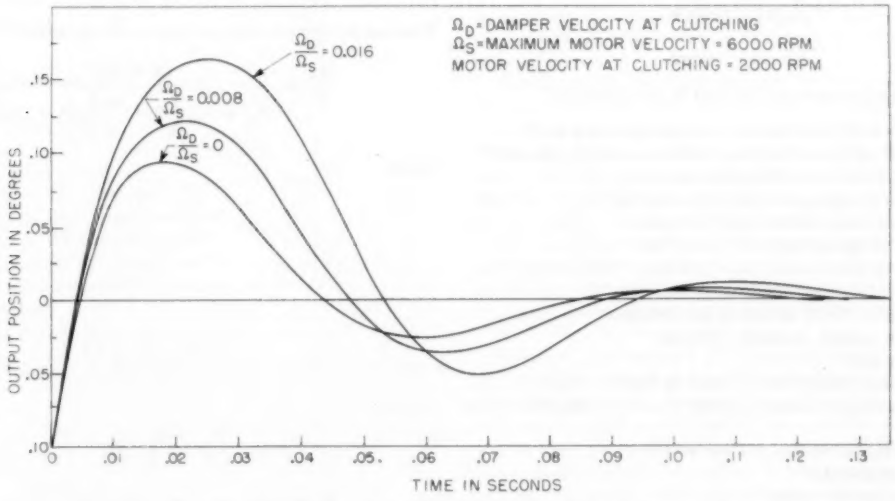


FIG. 13 EFFECT OF INITIAL DAMPER VELOCITY ON FINAL TRANSIENT

must be added the solenoid delay angle, equal to the maximum output rate (320 deg per sec) times the drop-out time of the solenoid (10 millisec), giving a total switching angle of 5.3 deg. The switching angle then represents the minimum error which will cause the system to switch to the high acceleration mode. With this setting, the system was subjected to step sizes ranging from 5.3 to 160 deg. The resultant reduction in settling time of the dual-mode system over the single-mode system is shown in Fig. 12. The analog-computer results show a higher factor of settling time reduction due primarily to an assumed ideal clutch characteristic.

As has been shown, the maximum torque transmission of the clutch affects the amount of overshoot to some degree. Further test results show that little is gained if the maximum clutch torque is raised above approximately four times the rated stall torque of the drive motor. This relaxes somewhat the requirements on the mechanical-clutch design. It is necessary for the clutch to exhibit very low drag torque in the disengaged state to prevent the damper from acquiring an appreciable velocity during slew. The overshoot magnitude and settling time increase measurably with increasing drag torque as observed in Fig. 13. If drag torque for a fixed clutch design is excessive, a brake may be used to stop the damper during slew.

[8] "Better Synchro Repeaters From Damper-Stabilized Feedback," by John E. Ward, *Control Engineering*, vol. 2, July, 1955, pp. 90–91.

[9] Velocity constant is defined as the ratio of the magnitude of a constant input velocity to the steady-state following error resulting from the constant velocity input.

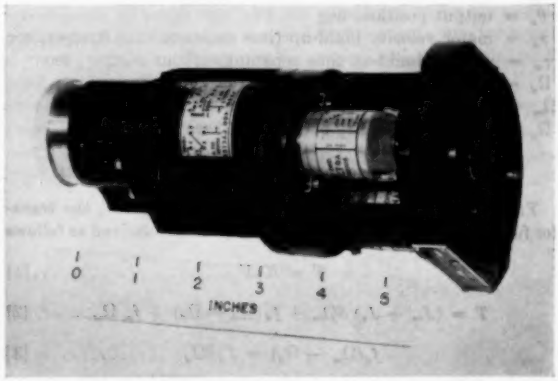


FIG. 14 SYNCHRO DATA REPEATER

nism showed negligible signs of wear and no deterioration in response characteristic after responding to 10,000 cycles of large-step operation.

Acknowledgment

The research reported in this paper was made possible through the support extended the Massachusetts Institute of Technology, Servomechanisms Laboratory, by the United States Air Force (Armament Laboratory, Wright Air Development Center), under Contract No. AF33(616)-2038, and by the joint support of the Department of the Army, the Department of the Navy, and the Department of the Air Force under Air Force Contract No. AF-19(122)-458 with Lincoln Laboratory, M.I.T., and executed under subcontract by the Servomechanisms Laboratory, M.I.T. It is published for technical information only and does not represent recommendations or conclusions of the sponsoring agencies.

Appendix

Nomenclature

The following nomenclature is used in the Appendix:

a_d = average early acceleration with damper, rad sec^{-2}
a_w = average early acceleration, without damper, rad sec^{-2}
f_d = damping constant of damper, in-oz sec rad^{-1}
f_m = internal damping of motor, in-oz sec rad^{-1}
n = gear ratio from motor shaft to output
J_d = inertia of damper slug, in-oz sec^2 rad^{-1}
J_m = inertia of motor rotor, gears and load, in-oz sec^2 rad^{-1}
J_s = inertia of damper shell, in-oz sec^2 rad^{-1}
K_1 = gear ratio^{-1} from output to fine synchro
K_2 = synchro output constant, volts deg^{-1}
K_3 = amplier gain
K_4 = conversion factor from radians to degrees, $180/\pi$
K_t = system torque constant, output torque per deg error, in-oz deg^{-1}
K_T = motor torque constant, in-oz volt^{-1}
S = Laplace operator
T = internal motor torque, in-oz
T_c = maximum slip torque of clutch, in-oz
T_s = stall torque of motor at rated control voltage, in-oz
V = applied motor voltage, volts
α = ratio of inertias defined by Equation [8]
β = ratio of inertias defined by Equation [14]
ϵ = error, deg
θ_i = input position, deg
θ_m = motor angular position, radians
θ_o = output position, deg
τ_d = major velocity build-up time constant with damper, sec
τ_w = velocity build-up time constant without damper, sec
Ω_d = damper angular velocity, rad sec^{-1}
Ω_m = motor angular velocity, rad sec^{-1}
Ω_s = maximum velocity of motor, rad sec^{-1}

Transfer Function for Damper-Stabilized System

From Fig. 15, using terminology previously given, the transfer function for the damper-stabilized system is derived as follows

$$T = K_T V \quad [1]$$

$$T = (J_m + J_s) S\Omega_m + f_d(\Omega_m - \Omega_d) + f_m \Omega_m \quad [2]$$

$$f_d(\Omega_m - \Omega_d) = J_d S\Omega_d \quad [3]$$

Eliminating Ω_d gives

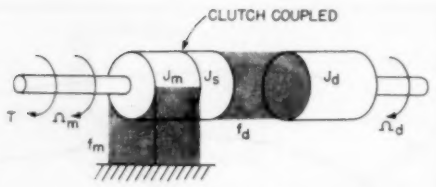


Fig. 15 Damper-Motor Schematic

$$\frac{\Omega_m}{T} = \frac{S + \frac{f_d}{J_d}}{(J_m + J_s)\left[S^2 + S\left(\frac{f_d}{J_m + J_s} + \frac{f_d}{J_d} + \frac{f_m}{J_m + J_s}\right) + \frac{f_m}{(J_m + J_s)}\frac{f_d}{J_d}\right]} \quad [4]$$

This can be written more simply in the approximate form

$$\frac{\Omega_m}{T} = \frac{(S + \omega_d)}{(J_m + J_s)\left(S + \frac{\omega_m}{\alpha}\right)(S + \alpha\omega_d)} \quad [5]$$

where

$$\omega_d = \frac{f_d}{J_d} \quad [6]$$

$$\omega_m = \frac{f_m}{J_m + J_s} \quad [7]$$

$$\alpha = \frac{J_m + J_s + J_d}{J_m + J_s} \quad [8]$$

Further

$$\frac{\theta_m}{V} = \frac{K_T}{S}\frac{\Omega_m}{T} \quad [9]$$

$$\frac{\theta_m}{V} = \frac{K_T (S + \omega_d)}{S(J_m + J_s)\left(S + \frac{\omega_m}{\alpha}\right)(S + \alpha\omega_d)} \quad [10]$$

The system is then shown in block diagram in Fig. 16.

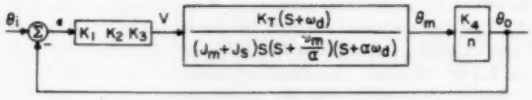


Fig. 16 System Block Diagram

For the damper declutched, the equations become

$$T = K_T V \quad [11]$$

$$T = J_m S\Omega_m + f_m \Omega_m \quad [12]$$

Hence

$$\frac{\theta_m}{V} = \frac{K_T}{J_m S(S + \beta\omega_m)} \quad [13]$$

where

$$\beta = \frac{J_m + J_s}{J_m} \quad [14]$$

For large step inputs the acceleration capabilities for the system with and without the damper are compared as:

(*a*) For the damper in the system

$$a_d = \frac{T_s}{\alpha\beta J_m} \quad [15]$$

$$\tau_d = \frac{\alpha}{\omega_n} \quad [16]$$

(*b*) For the damper disconnected

$$a_w = \frac{T_s}{J_m} \quad [17]$$

$$\tau_w = \frac{1}{\beta\omega_n} \quad [18]$$

Hence for the damper disconnected the acceleration is greater by the factor $\alpha\beta$ which, for the system considered, is about 25.

For the system described in this paper the constants have the following values:

$K_1 = 27$
$K_2 = 0.4$ volt deg^{-1}
$K_3 = 375$
$K_4 = 57.3$ deg rad^{-1}
$n = 108$
$K_T = 0.016$ in-oz volt^{-1}
$J_m = 0.72 \times 10^{-4}$ in-oz sec^2 rad^{-1}
$J_s = 1.41 \times 10^{-4}$ in-oz sec^2 rad^{-1}
$J_d = 1.69 \times 10^{-3}$ in-oz sec^2 rad^{-1}
$f_m = 1.23 \times 10^{-3}$ in-oz sec rad^{-1}
$f_d = 0.156$ in-oz sec rad^{-1}
$T_s = 2.4$ in-oz
$\Omega_s = 600$ rad sec^{-1}

Using these values, the transfer function of Equation [10] is found to be

$$\frac{\theta_m}{V} = \frac{75(S + 92.3)}{S(S + 0.65)(S + 825)} \quad [19]$$

$$\frac{K_1K_2K_3K_4}{n} = 2.15 \times 10^3 \text{ volt rad}^{-1} \quad [20]$$

$\tau_d = 1.5$ sec
$\tau_w = 0.0585$ sec
$K_t = 7000$ in-oz deg^{-1}

Discussion

G. A. Biernson.[10] Damper compensation is ideally suited for many instrument-servo applications because of the high stiffness and wide bandwidth that can be achieved with rather simple electronics. The dual-mode technique described in this paper provides an excellent solution to the poor synchronizing response which is the main limitation of damper compensation in such applications.

It should be pointed out, however, that when there is significant load inertia, damper compensation is generally inferior to tachometer compensation, because of the following:

(*a*) The sustained acceleration capability of the motor is much less when coupled to a damper, because the motor must accelerate the damper slug.

(*b*) For good stability the allowable reflected load inertia with respect to the motor must be quite small in comparison to the damper-slug inertia.

Consequently, the acceleration capability of the output is quite restricted with damper compensation if there is much load inertia.

The acceleration capability of a damper-stabilized motor is equal to the motor torque divided by the total inertia coupled to the motor shaft which includes the inertia of the floating slug and the direct-coupled inertia. When full motor voltage is first applied the damper-stabilized motor can achieve instantaneously a much faster acceleration because it need only accelera e the direct-coupled inertia. However, if it is to sustain an acceleration for a reasonable period of time, it also must accelerate the damper slug, which is usually at least six times as large as the direct-coupled inertia.

The direct-coupled inertia includes the motor-shaft inertia, the reflected load and gear inertia, and the damper-shell inertia. The damper-shell inertia, unfortunately, is from 5 to 10 per cent of the slug inertia and hence may account for as high as 60 per cent of the total direct-coupled inertia. Consequently the allowable value of the reflected load inertia is quite small in comparison to the total inertia coupled to the motor shaft.

The dynamic error of a damper servo in response to a low-frequency input is determined primarily by the acceleration-error coefficient. The value of this coefficient for the servo described in this paper unfortunately has not been mentioned.

Rufus Oldenburger.[11] The use of clutches as described here to modify the transfer function to improve performance is particularly effective for small servos, especially of the military variety where clutch wear may not be too big a problem. From our studies we have concluded that when it comes to large physical devices, such as governed engines, inertia-damping stabilization is out of the picture because of bulk, cost, and other factors. Further, our experience is that, when the clutch is slipping, the force that exists depends on a number of unknowns and cannot normally be taken into account in accurate analytical studies. If the clutching and declutching are done quickly enough this may not cause trouble.

Authors' Closure

We appreciate the time Mr. Biernson spent in preparing his comments although they are not considered to be appropriate. The purpose of the paper was to describe a technique for improving the response of a particular system and not to discuss the relative merits of various compensation schemes.

Also we do not consider it particularly unfortunate that the acceleration-error coefficient was not mentioned.

[10] Advanced Research Engineer, Sylvania Electric Products Inc., Electronic Systems Division, Waltham, Mass.

[11] Director of Research, Woodward Governor Company, Rockford, Ill. Mem. ASME.

Experiments With Optimalizing Controls Applied to Rapid Control of Engine Pressures With High-Amplitude Noise Signals

By GEORGE VASU,[1] CLEVELAND, OHIO

Optimalizer control principles were applied to the control of a flight-propulsion system. The control system described metered the fuel flow to an engine in such a manner as to cause the engine to seek a maximum pressure. Experimental data are presented illustrating the control behavior for a range of flight conditions, for various control settings such as gain and integral time constant, for various amounts of filtering, and so on.

Nomenclature

The following nomenclature is used in the paper:

- A_0 = amplitude of optimalizer test signal
- A_1 = amplitude of controlled pressure signal resulting from test signal
- A_n = amplitude of a particular noise-frequency component
- K = gain of proportional part of control
- $N(t)$ = general function denoting noise
- P_x = controlled pressure (engine-compressor output pressure), psfa
- t = time, sec
- V_1 = test-signal voltage, volts
- V_{mf} = filtered-multiplier output voltage, volts
- V_{mo} = unfiltered-multiplier output voltage, volts
- V_x = amplified voltage proportional to controlled pressure, volts
- V_{xf} = amplified and filtered signal at output of bandpass filter, volts
- W_f = fuel flow, lb/hr
- θ = phase angle between signals, radians
- τ = time constant of integral control, sec
- ω = frequency, radians/sec (rps)
- ω_n = frequency of a particular noise component, radians/sec (rps)

Introduction

In recent years, automation or automatic control has done much to stimulate progress. Of the many types of automatic controls possible, that class of controls employing optimalizing principles[2] is certain to become of increasing importance in the coming years as people become more familiar with such systems.

In the field of automatic controls for flight-propulsion systems there are numerous situations in which optimalizing principles can be applied. One of these situations is considered in this paper. Specifically, the paper presents a discussion of an optimalizing control which varies the input fuel flow to a flight-propulsion system in such a manner as to produce the maximum output pressure automatically. The paper first describes the type of optimalizer control used, with a discussion of its principle of operation, including the effect of component dynamics and noise on control operation. Next, the basic behavior of the system as determined experimentally is discussed; and finally, the effect on performance of filtering, test-signal amplitude, and so forth, is treated.

Theory of Control

Principle of Operation. For the application under discussion, the region of desirable engine operation can be related to the region near peaks in pressure fuel-flow characteristics such as shown in Fig. 1. The existence of such peaks suggests the possibility of utilizing optimalizer control principles[2] to insure operation in the desirable region throughout a range of flight conditions.

The particular type of optimalizer control chosen was the continuous test-signal type. A block diagram of the system is shown in Fig. 2. A sinusoidal test signal ($A_0 \sin \omega t$) was introduced at

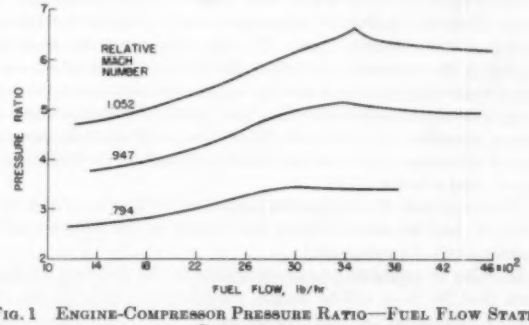


Fig. 1 Engine-Compressor Pressure Ratio—Fuel Flow Static Characteristics

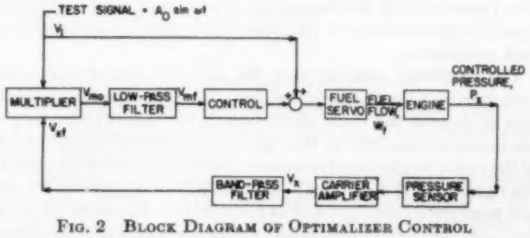


Fig. 2 Block Diagram of Optimalizer Control

[1] Assistant Head, Section A, Controls Branch, Lewis Flight Propulsion Laboratory, National Advisory Committee for Aeronautics.

[2] "Principles of Optimalizing Control Systems and an Application to the Internal Combustion Engine," by C. S. Draper and Y. T. Li, Aeronautical Engineering Department, Massachusetts Institute of Technology, published by ASME, September, 1951.

Contributed by the Instruments and Regulators Division of The American Society of Mechanical Engineers and presented at the ASME-AIEE Conference on Nonlinear Control Systems, Princeton, N. J., March 26–28, 1956.

Note: Statements and opinions advanced in papers are to be understood as individual expressions of their authors and not those of the Society. Manuscript received at ASME Headquarters, January 31, 1956. Paper No. 56—IRD-14.

the fuel servo input as illustrated. This test signal caused periodic variations in fuel flow and engine pressures. A particular pressure indicative of engine output was sensed and amplified. The resulting signal was then filtered with a bandpass filter to remove any noise present in the signal, to remove the steady-state or d-c pressure level, and to pass primarily the fundamental test-signal frequency.

The output of the bandpass filter consisted, then, essentially of a signal of test-signal frequency with an amplitude determined by the gains of the servo, engine, sensor, amplifier, and filter, plus some phase shift due to the dynamics of these same components. The signal at the multiplier (V_{xf}) thus can be represented by $A_1 \sin(\omega t + \theta)$.

The test signal is also introduced into the multiplier. The output of the multiplier is the product of the two inputs giving

$$V_1 V_{xf} = [A_0 \sin \omega t][A_1 \sin(\omega t + \theta)] \quad [1]$$

or

$$V_1 V_{xf} = A_0 A_1 \sin \omega t \sin(\omega t + \theta) \quad [2]$$

Utilizing the trigonometric formula

$$\sin \alpha \sin \beta = \tfrac{1}{2}[\cos(\alpha - \beta) - \cos(\alpha + \beta)] \quad [3]$$

gives

$$A_0 A_1 \sin \omega t \sin(\omega t + \theta) = \frac{A_0 A_1}{2}[\cos(-\theta) - \cos(2\omega t + \theta)] \quad [4]$$

or

$$V_{mo} = \frac{A_0 A_1}{2}[\cos(-\theta) - \cos(2\omega t + \theta)] \quad [5]$$

Thus the signal out of the multiplier consists of a d-c component and a second harmonic of the test-signal frequency.

The purpose of the low-pass filter following the multiplier is to pass the d-c component and to filter out the second harmonic. The control therefore receives essentially the d-c component.

The magnitude of the d-c component is a function of the amplitudes of the two multiplier input signals and is also a function of the phase shift between them. The $(A_0 A_1/2) \cos(-\theta)$ term of Equation [5] consists then of A_0 which is the amplitude of the test signal that was introduced directly into the multiplier; A_1 which is the test-signal amplitude times the gain of the servo, engine, sensor, amplifier, and bandpass filter; $\cos(-\theta)$ which is the cosine of the phase angle between the two signals at the multiplier input; and a factor of $\frac{1}{2}$.

The magnitude of the signal to the control is therefore $(A_0 A_1/2) \cos(-\theta)$ and the control action will depend on the factors which make up this d-c component.

In order to establish proper directions to the fuel-flow change such that the peak will be sought, consider the algebraic sign of the d-c component $(A_0 A_1/2) \cos(-\theta)$. A_0 and A_1 are amplitudes of periodic signals and can be thought of as having positive signs. The phase angle θ will depend on dynamics of components and upon the algebraic sign or the slope of the static characteristics of each component.

For simplicity, consider that the algebraic sign of each component except the engine is positive. In the engine the algebraic sign or slope of the static characteristic is positive to the left of the peaks and negative to the right. In other words, to the left of the peaks (neglecting dynamics) the phase angle θ is zero and to the right it is 180 deg. $\cos(-\theta)$ then will be positive on the left and negative on the right and as a result the whole term $(A_0 A_1/2) \cos(-\theta)$ (neglecting dynamics) will be positive when the engine is operating to the left of the peaks and negative to the right. When operating on the left, the control then receives a positive signal which is fed to the fuel servo to increase fuel flow. When operating on the right of the peaks, the signal is reversed, causing the fuel flow to decrease. In either case, the control will seek the peaks in the characteristics and the system will operate in the desirable region.

Effect of Component Dynamics. Thus far, component dynamics have been neglected. The effect of dynamics is to introduce phase shift in the signals traveling through the system and also to attenuate or amplify the various frequency components, so that usually the gain for a particular frequency is different from the zero-frequency gain associated with the static characteristics.

The effect of dynamics on the control action resulting from the $(A_0 A_1/2) \cos(-\theta)$ term of Equation [5] can be defined as follows: (1) Any attenuation or amplification of the test signal frequency as it passes through the components from the servo to the multiplier will affect A_1, and A_1 should be determined by the gains of components at the test-signal frequency if there are appreciable dynamics in these components. (2) The output of the multiplier will be affected as a function of $\cos(-\theta)$ and the signal to the control will vary accordingly.

Consider the variation of $\cos(-\theta)$ as a function of θ. For $0 < \theta < 90$ deg, $\cos(-\theta)$ decreases from 1 to 0. For $90 < \theta < 180$ deg, $\cos(-\theta)$ varies from 0 to -1. Therefore phase shifts of up to 90 deg cause an attenuation of the signal to the control, but polarity is still suitable for satisfactory behavior. Phase shifts of more than 90 deg, however, cause a change in sign to the control which would cause divergent or runaway control action.

A method of eliminating the problems arising from phase shift caused by component dynamics is to insert an equal amount of phase shift into the test signal before it enters the multiplier. The two signals entering the multiplier would then be in proper phase relation.

Effect of Noise. With noise present in the system, the signal to the multiplier from the bandpass filter can be considered as $A_1 \sin(\omega t + \theta) + N(t)$. The effect of the first term $A_1 \sin(\omega t + \theta)$ which is the test signal frequency component has been discussed. The noise $N(t)$, in general, will be made up of a large number of frequency components. Consider a typical frequency component $A_n \sin \omega_n t$. The contribution to the multiplier output due to this particular component of noise will be

$$A_0 A_n \sin \omega t \sin \omega_n t = \frac{A_0 A_n}{2}[\cos(\omega t - \omega_n t) - \cos(\omega t + \omega_n t)] \quad [6]$$

These are the familiar sum and difference frequencies arising in a modulation or multiplication process. No d-c component is produced unless a noise component exists with exactly the same frequency as the test signal. In many cases such an occurrence is unlikely. When such an occurrence is likely, proper attention must be given to the choice of the test-signal frequency, to filtering action, and to the amplitude of the test signal used.

The sum and difference frequency components in the multiplier output would generally be of small magnitude as a result of the filtering action of the bandpass filter on the noise in the controlled pressure signal. In any case, whether they are large or small, they would contribute to control action and the result would be a control response with these superimposed frequency components. As long as the noise frequency is different from the test-signal frequency, the product will be an a-c signal which will produce no change in the average value of fuel flow. If the magnitude of the a-c components is objectionable, the solution is to filter out those frequencies.

It appears, then, that the multiplier does an excellent job of dis-

criminating between noise and the test signal. There is a special situation which may cause some difficulty, however. If a strong noise component exists at a frequency close to that of the test signal, the difference frequency component produced in the multiplication process would be a very low-frequency signal, causing the fuel flow to vary at this frequency. The test-signal frequency should be chosen so that there is no single noise component with a large amplitude and a frequency near that of the test signal. Large amplitudes of noise at frequencies farther from the test-signal frequency can, of course, be filtered out as previously mentioned.

Description of Experimental System

Pressure Sensor. The engine pressure being controlled was sensed by a variable-inductance type of pressure transducer. The tubing connecting the transducer to the pressure probe in the engine was made as short as possible to reduce dynamic effects. The frequency response of the sensor with tubing was approximately equivalent to that of a second-order system with a damping ratio of 0.5 and a natural frequency of 130 cycles per second (cps). At 20 cps there was less than 5 per cent rise in amplitude ratio and less than 10 deg phase shift. At frequencies below 20 cps, which is the range of most significance in the control system, amplitude and phase errors were very small and can be neglected.

Carrier Amplifier. The output of the sensor was amplified by a carrier-type amplifier designed for frequencies from zero to 300 cps; therefore, for frequencies below 20 cps, amplitude and phase errors were again very small.

The gain of the sensor and carrier amplifier was 1.08×10^{-3} volts per psf.

Bandpass Filter. The bandpass filter used in the investigation was a commercial type with adjustable upper and lower cutoff frequencies. The gain in the pass band was approximately unity. On each side of the pass band the attenuation was 24 decibels per octave. The transfer function of the filter was approximately

$$\frac{\omega^4\tau_1{}^4}{(1 + 2jA\omega\tau_1 - \omega^2\tau_1{}^2)^2(1 + 2jA\omega\tau_2 - \omega^2\tau_2{}^2)^2}$$

where the low cutoff frequency is $1/2\pi\tau_1$, the high cutoff frequency is $1/2\pi\tau_2$, and the factor A is slightly greater than 0.6.

For all of the data presented except Figs. 3 and 4 and where specifically indicated, the low cutoff frequency was 1 cps and the high cutoff frequency was 4 cps.

For Figs. 3 and 4 the low cutoff frequency was 0.5 cps and the high cutoff frequency was 8 cps.

D-C Amplifiers. An electronic analog computer was used for that part of the control system from the bandpass-filter output to the fuel servo input. The output of the bandpass filter was amplified by a d-c amplifier in the computer before the multiplication was performed. The gain of this amplifier was 500.

The test signal V_i, generated by a low-frequency function generator, was also amplified before being fed to the multiplier. The gain of this amplifier was 10.

For both d-c amplifiers the bandwidth was sufficient to permit dynamic errors below 20 cps to be neglected. In the block diagram of Fig. 2 these amplifiers have been omitted for simplicity.

Multiplier. The multiplier was a pulse-width, pulse-height modulation type of multiplier with a bandwidth of approximately 1000 cps and therefore produced negligible amplitude and phase error at pertinent control frequencies. The multiplier accepts two inputs x and y and produces a product $z = xy/100$.

Low-Pass Filter. The low-pass filter consisted of a simple lag with the transfer function

$$\frac{K_1}{1 + \tau_2 p}$$

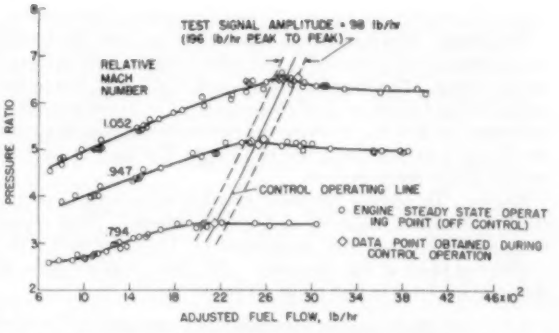


Fig. 3 Control Operating Line on Engine Pressure-Fuel Flow Characteristics

(Test-signal amplitude, $A_0 = 1.41$ volts; control gain, $K = 0.05$; integrator time constant, $\tau = 0.05$ sec.)

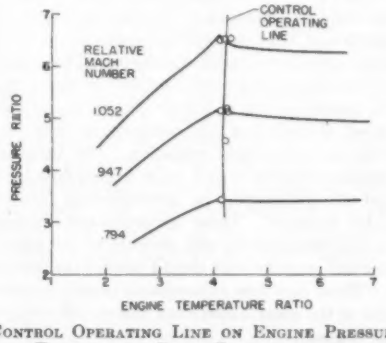


Fig. 4 Control Operating Line on Engine Pressure Ratio-Temperature Ratio Characteristics

(Test-signal amplitude, $A_0 = 1.41$ volts; control gain, $K = 0.05$; integrator time constant, $\tau = 0.05$ sec.)

where the gain K_1 was unity and the time constant τ_2 was 0.1 sec which gives a break frequency of approximately 1.6 cps.

Control. A proportional-plus-integral control was chosen for the investigation; thus additional flexibility over a simple proportional control or integral control was obtained. Also, experience has indicated that this type of control is suitable for a wide variety of systems. For this particular system where the predominant dynamics are in the filters, a proportional-plus-integral control appears to be especially desirable.

The transfer function of the control was

$$K\left(1 + \frac{1}{\tau p}\right)$$

where the control gain K and the integrator time constant τ were variable so that the control performance could be evaluated for different control settings.

During a portion of the investigation, the proportional part of the control was disconnected to give plain integral control. The transfer function in this case reduced to

$$K/\tau p$$

where the effective control time constant is τ/K.

All of the data presented herein are for proportional-plus-integral-control action except those in Fig. 8 which were obtained with integral control only.

Summer. The output of the control was added to the test signal as shown in Fig. 2. The gain from the control output to the fuel servoinput was unity, but the test signal was attenuated by a

factor of 3, resulting in a gain of $^1/_3$ for that portion of the summer.

Fuel Servo. The fuel system consisted of an electrohydraulic servosystem which positioned a fuel valve in response to an input-voltage signal. The frequency response of fuel flow to fuel servo-input voltage for the range of frequencies up to 20 cps is roughly equivalent to that of a second-order system with a damping ratio of 0.5 and a natural frequency of 10 cps. The gain of the fuel servo is 278 lb per hr per volt.

Engine. To provide favorable conditions for control operations, a test-signal frequency of 2 cps was selected. This choice of frequency was made for two reasons: (1) The phase shift in the engine was small at that frequency. (2) There were no peaks in the noise spectrum in that range of frequencies.

Behavior of Experimental System

Steady-State Performance. The control system described was operated at various simulated flight conditions. The steady-state performance attained over a range of flight conditions is illustrated in Fig. 3. The figure shows engine operation in terms of compressor-pressure ratio as a function of adjusted fuel flow. For a particular altitude and Mach number the ordinate can be considered as the engine output or controlled pressure. The data have been plotted as a function of adjusted fuel flow to eliminate, as much as possible, variations due to burner efficiency and variations in Mach number, altitude and inlet temperature from the intended values at the simulated flight condition. The adjustments were made by multiplying the actual value of fuel flow by burner efficiency and also multiplying by generalization factors for temperature and pressure. These generalization factors involve ratios of actual temperature and pressure to the exact values of temperature and pressure that should exist at the particular flight condition. These fuel-flow adjustments permit a more accurate presentation of the control operating line on the static characteristics of the engine.

The steady-state values of pressure and fuel flow obtained during control operation are indicated in the figure. The control gain K was 0.05; the integrator time constant τ was 0.05 sec; and the test-signal amplitude A_0 was 1.41 volts. The test-signal amplitude is indicated in the figure in terms of fuel-flow to show the magnitude of cyclic perturbations.

The data indicate that very nearly the peak pressure was attained at all Mach numbers. Even at the lowest Mach number, where the slope to the right of the peak is very small, the system operated near the peak. It appears that the control is causing the engine to operate slightly to the right of the peaks over the range of Mach numbers. The amount of deviation from the peak in most cases is quite small and is within the range of the experimental error involved in determining the static characteristics.

To establish the control operating line more definitely and permit easier visualization, the same steady-state performance data shown in Fig. 3 are presented in Fig. 4 in terms of compressor-pressure ratio and engine-temperature ratio. Since ratios are involved on both ordinate and abscissa, the engine characteristics can be fixed more accurately. Also, the peaks of the engine characteristics fall at nearly the same temperature ratio, resulting in a nearly vertical control operating line. Again the data points obtained during control operation fall slightly to the right of the peaks.

There are at least two possible explanations for the position of the control line. First, any static error caused, for example, by drift in certain components such as the multiplier, would result in operation off the peak, and could cause some of the discrepancies shown. The second factor which certainly should cause operation to the right of the peaks is the result of different slopes below and above the peaks. The steeper slope to the left would give a signal during the cyclical excursion to the left which is stronger than that obtained on the right side. As a result, the system should settle out slightly to the right of the peaks.

Another factor affects the value of pressure obtained during control operation. For test-signal excursions around the peak, the cyclic pressure resulting would have a lower average value than the peak of the static characteristics. This hunting loss is quite small, however.

Transient Performance. The dynamic performance of the system was studied by subjecting the control system to step disturbances in fuel servo input voltage. A typical response of the control system to such a disturbance is shown in Fig. 5. The test signal amplitude A_0 was 2.83 volts (262 lb/hr in fuel flow); the control gain K was 0.05; and the integrator time constant τ was 0.2 sec. The control was operating in steady state when, at point A, the step disturbance was introduced causing the servo input voltage to change suddenly to a new value (point B). As a result of the step change in servo input voltage the fuel-valve position and fuel flow also change rapidly to new values. Notice that just before the transient, there is a gap in the fuel-flow trace.

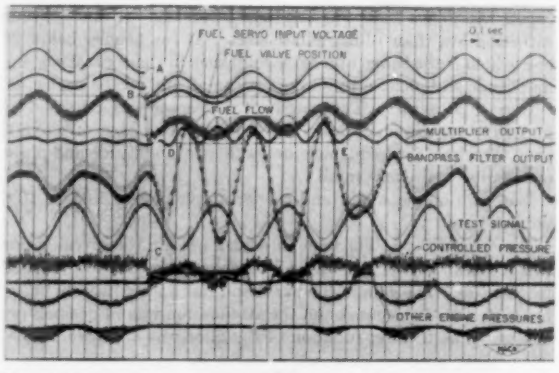


Fig. 5 Response of Control System to Step Disturbance in Fuel Servo Input

(Relative Mach number = 0.847; step size = 2v [556 lb/hr]; test-signal amplitude, A_0 = 2.83 volts; control gain, K = 0.05; integrator time constant, τ = 0.2 sec.)

This gap is caused by the recorder-channel identification system and coincidentally occurred at the same time as the transient was introduced. A very short time after the fuel flow changed, the controlled pressure responded to the disturbance (point C). Since the step disturbance was a step decrease in fuel flow, the controlled pressure dropped, and then responded periodically and approximately in phase with the fuel flow. The amplitude of the disturbance (556 lb/hr or 18 per cent of the steady-state value) was roughly equal to the peak-to-peak test-signal amplitude (524 lb/hr). The test-signal trace happens to be reversed in phase on the recording so that a deflection downward on the trace corresponds to deflections upward on the servo input voltage, fuel-valve position, fuel flow, and controlled-pressure traces.

The controlled-pressure signal was filtered with a bandpass filter, the output of which is shown directly above the test signal. Several cycles of test-signal frequency are clearly evident during the time interval from D to E just following the step input.

The next trace above the bandpass-filter output is the multiplier output which is the product of the test signal and the bandpass filter output. The double frequency or second harmonic produced as a result of the multiplication is also clearly evident in the time interval from D to E. In order to show the d-c component, a line has been drawn indicating the average steady-state value of the multiplier output before and after the transient. During the D to E time interval, the minimum value of the multi-

plier output signal falls along this line, indicating that the d-c component is one half the peak-to-peak value of the amplitude of the second harmonic signal produced by the multiplier. This d-c component is essentially constant in this interval and the integral action of the control causes the fuel flow to increase at a constant rate up to point E. At that time the d-c level decreases as the system approaches the final value. In steady state, the multiplier output oscillates plus and minus a small amount as a result of excursions in fuel flow to the right and to the left of the peak. The oscillations in controlled pressure resulting from the test signal are barely discernible through the noise present in the engine during steady-state operation.

Effect of Control Settings on Dynamic Performance. The control system was subjected to a series of disturbances similar to the one just discussed. For each of these transients, different control settings were used. The responses were then analyzed to determine the response time and per cent overshoot. These were measured on the fuel servo input-voltage trace to facilitate measurements and to avoid difficulties in interpretation resulting from pressure reversals in passing through the peaks of the engine characteristics. A line was drawn through the fuel servo input oscillations to indicate the average value of the oscillations during the transient. This average value was used for the calculations of response time and per cent overshoot. Response time was defined as the time from the disturbance until the transient first completed 90 per cent of its response. Per cent overshoot was defined as the ratio of the first overshoot to the disturbance magnitude.

Response time and per cent overshoot are shown as a function of integrator time constant for a fixed value of control gain (K = 0.1) and fixed disturbance magnitude (556 lb/hr) in Fig. 6. The test-signal amplitude A_0 was 2.83 volts. The response time increases and per cent overshoot decreases as integrator time constant increases. This trend is as expected. The response times for step increases are larger than for step decreases. This trend is caused by the less steep slope to the right of the peak than that to the left.

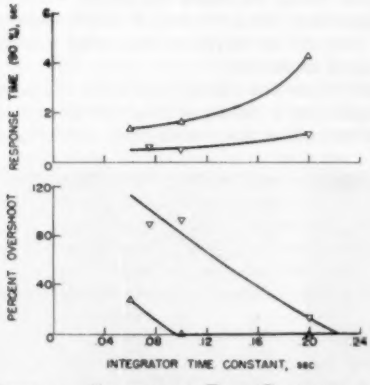


FIG. 6 EFFECT OF INTEGRATOR TIME CONSTANT ON CONTROL SYSTEM RESPONSE

[Relative Mach number = 0.847; step size = 2v (556 lb/hr); test-signal amplitude, A_0 = 2.83 volts; control gain, K = 0.1.]

The relation between per cent overshoot and response time is shown in Fig. 7. The figure indicates that the minimum response time without overshoot is approximately 1.6 sec for step increases and 1.3 sec for step decreases at a relative Mach number of 0.847. Per cent overshoot increases rapidly below these values of response time.

Effect of Shape of Static Characteristics on Dynamic Performance. The shape of the static characteristics varies considerably over the range of simultated flight conditions as shown in Fig. 1. At high Mach numbers the characteristics tend to curve upward in the region near the peak, while at low Mach numbers, the characteristics tend to be rounded off near the peak, with a very small slope to the right.

A series of responses was obtained at the different Mach numbers with constant control settings and magnitude of disturbance. The test-signal amplitude A_0 was 2.83 volts. Also, the proportional part of the control was disconnected, thus the control action was integral only. Its effective time constant τ/K was 4 sec. For all of the other data presented, the control was of the proportional-plus-integral type. Response times and per cent overshoot for this series are shown in Fig. 8. The response times for step increases are considerably larger than for the step decreases, and for either direction of the step the response time decreases as the Mach number increases. Increasing Mach number corresponds to moving toward the cusp-type peaks.

There was no overshoot for any of the responses except the response to a step decrease at the highest Mach number.

The reasons for the differences in response become more evident from a consideration of Fig. 9. This figure shows the engine gain or slope of the pressure-fuel flow static characteristics as a function of fuel flow. Starting at low fuel flow, the engine gain is about the same magnitude and increases at about the same rate up to about 2400 lb per hr for all flight conditions. From 2400 lb per hr to the peak, the engine gain varies considerably for the different Mach numbers. At high Mach numbers, the gain continues to increase until the peak of the static characteristic is reached, and then changes abruptly to a large negative value.

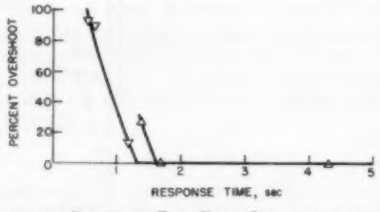


FIG. 7 RELATION BETWEEN PER CENT OVERSHOOT AND RESPONSE TIME

(Relative Mach number = 0.847; step size = 556 lb/hr; test-signal amplitude, A_0 = 2.83 volts; control gain, K = 0.1.)

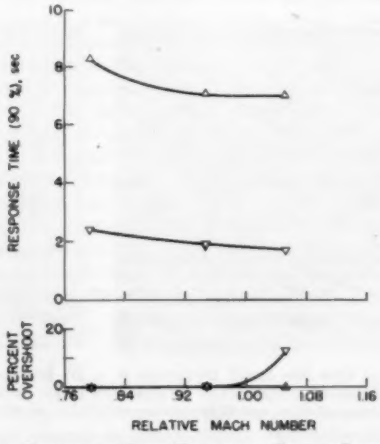


FIG. 8 EFFECT OF MACH NUMBER ON SYSTEM RESPONSE

[Step size = 2v (556 lb/hr); test-signal amplitude, A_0 = 2.83 volts; integral control only; effective integrator time constant, τ/K = 4 sec.

A short distance to the right of the peak the gain changes rapidly and approaches a small negative value.

For the lowest Mach number, the slope decreases before the peak is reached, goes through zero at the peak, then levels off at a very low negative value.

The signal, or error, to the control is proportional to the engine gain. The engine gain and hence this error is larger at higher Mach numbers, therefore, response times are shorter. Also, the error signal available on the right of the peaks is much less than that on the left, resulting in much slower responses for step increases than for step decreases. The response times for step increases can be made approximately the same as for step decreases by properly biasing the multiplier output.

The abrupt change in engine gain from positive to negative at the peaks of the high Mach number engine characteristics results in control operation similar to that obtained with relay or on-off type of controls. At the lowest Mach number the engine gain changes gradually through zero giving control performance more closely related to that in a linear continuous system for small disturbances to the left of the peak.

Effect of Filtering on Performance. The effect of filtering was investigated to determine how sensitive this type of optimalizer control was to the large amount of noise present in the controlled pressure signal.

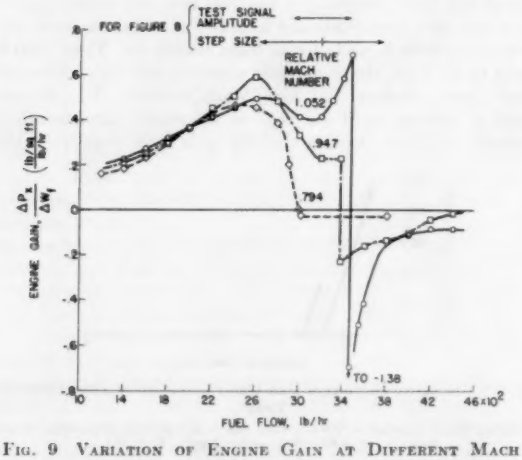


FIG. 9 VARIATION OF ENGINE GAIN AT DIFFERENT MACH NUMBERS

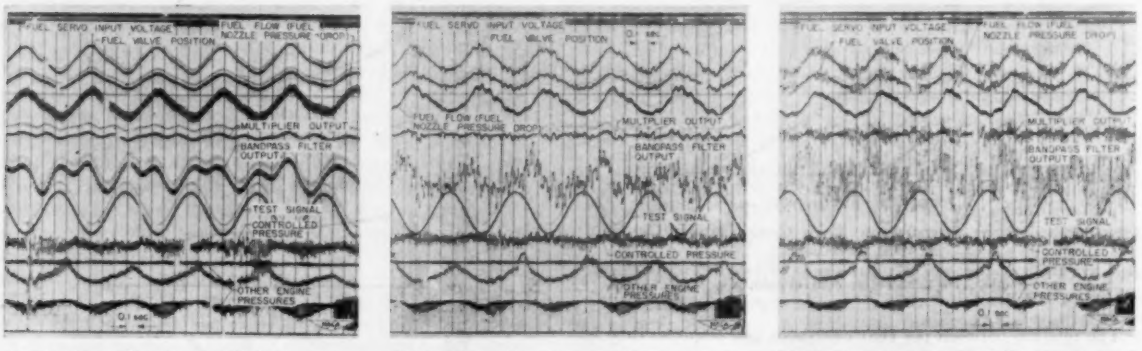
The low-pass filter following the multiplier was removed and the bandpass filter upper cutoff frequency was varied. The steady-state response of the system is shown in Fig. 10 for various cutoff frequencies. In Fig. 10(*a*) the test-signal frequency was approximately 2 cps, the bandpass-filter lower cut-off frequency was 1 cps, and the upper cutoff frequency was 4 cps. These were the original settings which were used during most of the investigation. In Fig. 10(*b*), the upper cutoff frequency was raised to 40 cps. The higher frequency components are apparent in the bandpass-filter output, multiplier output, fuel servo input, fuel-valve position, and fuel flow.

When the upper cutoff frequency was raised to 400 cps, the amount of noise passing through the system was considerably more, as indicated in Fig. 10(*c*). In spite of the very large noise level, there was no indication of unsatisfactory performance.

Performance for Various Test-Signal Amplitudes. In view of the large noise level present in the pressures in the engine, there was some concern as to the feasibility of operating with a reasonably small test-signal amplitude. During the course of the study this amplitude was varied in an attempt to find a lower limiting value of signal amplitude that was suitable in the presence of the noise. It was found, however, that the noise in the system was not the factor which determined the minimum allowable test-signal amplitude. Difficulties were encountered during transients at the lowest test-signal amplitude. These difficulties were apparently caused by small imperfections in the pressure fuel-flow static characteristics. Data have indicated that in many cases small deviations or peaks occur in regions on both sides of the main peak. These usually can be attributed to pressure-profile effects, and many times are not apparent from ordinary plots of steady-state data points. When such peaks have been found, they have been obtained by recording pressure as a function of fuel flow in a continuous plot on an *X-Y* recorder.

Although the system operated satisfactorily in steady state when the test-signal frequency was not apparent in the controlled pressure signal, during transients, the system would hang up or seek what apparently was a minor peak in the engine characteristic. This difficulty no longer occurred after returning to the larger test-signal amplitudes.

Behavior for Higher Test-Signal Frequencies. To insure successful control operation, a test-signal frequency of approximately 2 cps was used for most of the investigation. Since higher test-sig-



(*a*) Bandpass filter low cutoff frequency = 1 cps; bandpass filter high cutoff frequency = 4 cps; low-pass filter cutoff frequency = 1.6 cps

(*b*) Bandpass filter low cutoff frequency = 1 cps; bandpass filter high cutoff frequency = 40 cps; low-pass filter removed

(*c*) Bandpass filter low cutoff frequency = 1 cps; bandpass filter high cutoff frequency = 400 cps; low-pass filter removed

FIG. 10 EFFECT OF CHANGING FILTER CUTOFF FREQUENCY

(Relative Mach number = 0.847; test-signal frequency = 2 cps; test-signal amplitude, A_0 = 2.82 volts (262 lb/hr); control gain K = 0.1; integrator time constant, τ = 0.1 sec.)

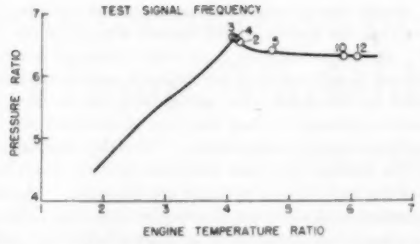


FIG. 11 STEADY-STATE PERFORMANCE OF CONTROL FOR VARIOUS TEST-SIGNAL FREQUENCIES

(Test signal amplitude A_s = 1.41 volts; control gain, K = 0.05; integrator time constant, τ = 0.1 sec; relative Mach number = 1.052.)

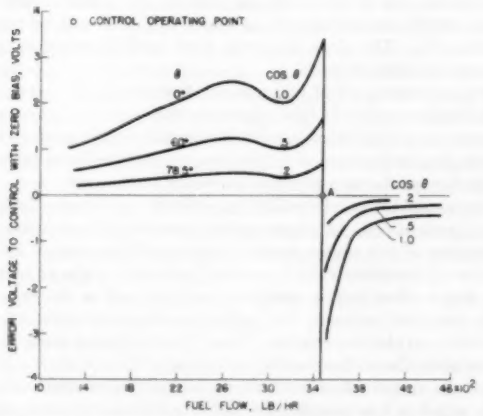


FIG. 12 EFFECT OF PHASE SHIFT ON ERROR VOLTAGE

(Test-signal amplitude, A_s = 1.41 volts; relative Mach number = 1.052.)

nal frequencies would permit higher filter cutoff frequencies and hence, faster response, data were taken to determine the performance of the system under such conditions.

The steady-state behavior of the system at several test-signal frequencies is shown in Fig. 11. No attempt had been made to correct for the phase shift due to the fuel servo and engine at these higher frequencies. Nevertheless, the system operated satisfactorily in steady state for frequencies up to 4 cps.

The error voltage to the control for these frequencies is indicated in Fig. 12. As the phase shift increases from 0 to 90 deg, the error voltage to the control $(A_0A_1/2)\cos\theta$ is reduced, and for 90-deg phase shift becomes zero. Operation at 90-deg phase shift therefore is impossible. For frequencies which produce less than 90-deg phase shift, the control operates at point A, which corresponds to operation at peak pressure.

When the frequency was changed from 4 to 5 cps, the operating point suddenly shifted from the peak to a point well to the right of the peak as indicated in Fig. 11. This sudden change in operating point occurred because the phase shift increased beyond 90 deg. For phase shifts greater than 90 deg (and less than 270 deg), $\cos\theta$ is negative. When $\cos\theta$ is negative, the control is unstable (runs away) at the peaks.

The conditions under which the system can operate stably far to the right of the peaks are illustrated in Fig. 13, where the error voltage to the control $(A_0A_1/2)\cos\theta$ is shown again. In this case, however, the curves are inverted because $\cos\theta$ is negative. Stable operation at high fuel flows is possible if a small amount of d-c bias exists (caused for example by drift in the multiplier).

The operating point shown in Fig. 11 for a test-signal frequency

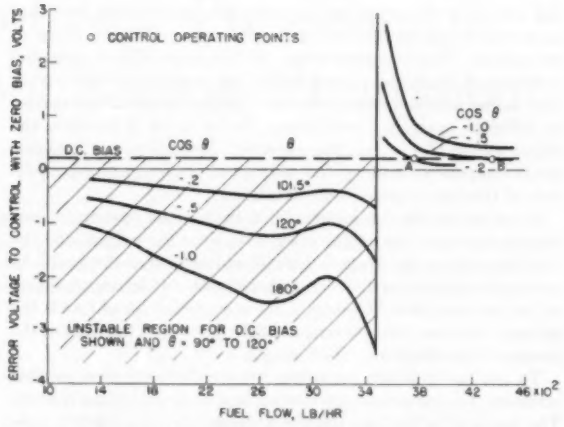


FIG. 13 EFFECT OF PHASE SHIFT ON ENGINE OPERATING POINT FOR ABNORMAL OPERATING CONDITIONS

of 5 cps corresponds to operation at point A in Fig. 13. Increasing the test-signal frequency from 5 to 12 cps caused the phase shift to increase further. As the phase angle increased, $\cos\theta$ increased in magnitude. The increase in error voltage to the control as a function of $\cos\theta$ caused the operating point to shift as illustrated by points A and B in Fig. 13 and as shown in Fig. 11 for frequencies from 5 to 12 cps.

The effect of component dynamics on A_1 in the term $(A_0A_1/2)\cos\theta$ must of course also be considered. In this case, however, variations in A_1 are small compared to variations in $\cos\theta$.

To alleviate the difficulties encountered here, and possibly permit operation at even higher test-signal frequencies, the test signal fed to the multiplier should be shifted in phase to compensate for the phase shift through the fuel servo, engine, and other components in the loop.

CONCLUSIONS

The experimental data indicated that the continuous test-signal optimalizer control, applied to the control of compressor-output pressure in a flight-propulsion system, gave very nearly maximum output pressure over a range of flight conditions.

For a test-signal frequency of 2 cps and with a large amount of filtering in the system, response times of less than 2 sec were obtained with only a small amount of overshoot. These response times can be shortened considerably by increasing the test-signal frequency and decreasing the amount of filtering.

The type of response varied greatly over the range of Mach numbers owing to the variation in the shape of the engine static characteristics from one Mach number to another.

The data indicated that good filtering was not required. The system could therefore be made to operate faster by removing part or all of the filtering.

The minimum test-signal amplitude was found to be dictated not by the signal-to-noise ratio, but by imperfections or minor peaks in the engine static characteristics. For very small test-signal amplitudes, the control would hang up on a minor peak.

Discussion

Y. T. LI.[3] The author is to be congratulated on this excellent paper. The presentation of the problem is well done and the experimental coverage is quite complete. An optimalizing system,

[3] Department of Aeronautical Engineering, Massachusetts Institute of Technology, Cambridge, Mass.

like any control system, must prove its practicability by being subjected to the disturbance environment the system is likely to encounter. The extensive study of the noise effects upon the operation of the tested system fulfills this requirement quite well. The author also has established the response speeds of the system at different operating conditions. To be more inquisitive, the writer would like to ask this question: How does this response speed compare with the actual drift speed of the operating condition of the engine under flight conditions?

In so far as the basic purpose of the control system is concerned, the use of the engine pressure ratio as the output, and the fuel flow rate as the controlled input, indicates that the goal is to maintain a maximum delivery of thrust without the consideration of fuel consumption. Clearly, if the economical use of fuel is the primary purpose, then the output might have been the ratio of the pressure ratio divided by the fuel rate.

To get an optimum operation we usually have two possible schemes; i.e., program-type control and an optimalizing control. The final choice between the two depends upon the relative complexity of the system. The writer wishes to know whether the author has considered the possibility of getting similar control performance with a programmed-type controller, and whether it is more complicated to attempt this?

In so far as an optimalizing controller is concerned, a continuous test signal is a neat system and has a lot of appeal to electronic systems operating at many thousands of cycles per second. For mechanical systems with hunting periods down to the order of a second, it is the opinion of the writer that a peak-holding type might be considerably more simple and flexible.

Author's Closure

The author appreciates the complimentary remarks and excellent discussion offered by Dr. Li. It is a privilege to have one of the inventors of optimalizing controls discuss this paper.

Dr. Li asks several rather interesting questions. In his first question he asks how the response speeds of the system, as investigated, compare with the actual drift speed of the operating condition of the engine under flight conditions.

The answer to this question will depend upon the particular application of the engine. For certain vehicles and missions, the response times given are adequate. For others, the speed of response is not so fast as desired. The response speeds shown, however, are for the system with a large amount of filtering. It was shown that the system would operate with all of the filtering removed. Since the response time would improve considerably as the filtering is reduced, it is the author's opinion that response times could be obtained with an optimalizing control system which would approach the best that can be obtained with standard nonoptimalizing control systems. Once the filtering has been removed, the author feels that the basic limitations on speed of response of the two types of systems are the same. (Actually, it is not suggested that all of the filtering be removed. How much filtering exists, either intentional or unavoidable, will again depend upon the specific application and on the hardware employed. Nevertheless, it is felt that the response times shown can be considerably improved.)

The second question involves the variable being optimalized. Apparently, the author's use of pressure ratio was confusing. The quantity actually sensed, as shown in Fig. 2, was an engine pressure, P_s. The plots, however, were made in terms of pressure ratio for convenience.

The economical use of fuel was not discussed in the paper and Dr. Li's observations in this regard are well made. It turns out, however, that minimum specific fuel consumption occurs at or near maximum pressure and therefore use of the ratio of pressure to fuel flow as the output is not necessary in this case.

Programmed-type controllers have been used successfully in many applications. A program-type control depends upon the calibration of the engine to get the required program. As the number of variables (Mach number, altitude, angle of attack, etc.) which affect engine operation increases and as the range of these variables increases, the calibration becomes more difficult to obtain and also less reliable. In addition, changes occur in the engine after the calibration has been made.

It is the author's opinion that an optimalizing control could be made which is less complicated and more reliable than many of the existing engine controls. Operations such as filtering, multiplication, and signal generation can be performed by relatively simple devices which can be mechanical, electrical, electronic, acoustic, or other types. Utilizing such devices, simple continuous test-signal optimalizing controllers can be designed to operate over a wide range of frequencies. Although the peak-holding type might be more simple and flexible, it would very likely not be so insensitive to noise as the continuous test-signal type.

Representation of Nonlinear Functions of Two Input Variables on Analog Equipment

By D. A. ELLIOTT,[1] CALDWELL, N. J.

When an analog operator who has been solving problems on a linearized basis begins to employ nonlinear components to represent a wide range of operation of the simulated system, it soon becomes apparent that many of the key relations in the system are nonlinear functions of two or more input variables. Function-generating equipment that will develop nonlinear functions of a single input variable is now available for many types of analog equipment, but equipment that will generate directly three-dimensional families of curves as arbitrary nonlinear functions of two input variables exists only in a few cases of highly expensive specialized equipment. Some methods also have been described for developing three-dimensional curves by reworking standard function-generating equipment to perform the additional mathematical operations required by a second input variable. This paper describes a graphical method of matching three-dimensional functions by the use of standard analog components without modifications.

Introduction

THE method which is described here uses standard analog components without any modifications for accomplishing the simulation of three-dimensional functions. A graphical method of matching the desired family of curves is used to determine the necessary analog relations. This provides a simple method having reasonably good accuracy, which is readily adaptable to a large variety of functions. Although the method depends upon a somewhat regular progression among the lines in a family of curves, most functions do have such a progression. With this method it has been possible to match all curves which have been encountered in the simulation of several types of ramjet and turbojet engines with sufficient accuracy to achieve the aim of analog computation; i.e., the study of dynamic-stability characteristics in control-analysis problems. For such studies, the principal requirement is that the correct slopes, or partial derivatives, be present to relate the variables at any particular operating point.

Basic Procedures

In order to describe this method, it is first necessary to define the terms to be used. The familiar procedure of using delta increments is useful for nonlinear functions in a manner similar to that of linear-analog systems, where voltages represent excursions about a steady-state point. When used with nonlinear systems this permits greater accuracy, since the range of variation of each variable can be made to cover the full range of variation of available voltage. An illustration of the delta method as applied to a fixed coefficient is provided by Fig. 1. This relation is valid for either linear or nonlinear analog representations, and serves to define the terms which will be used.

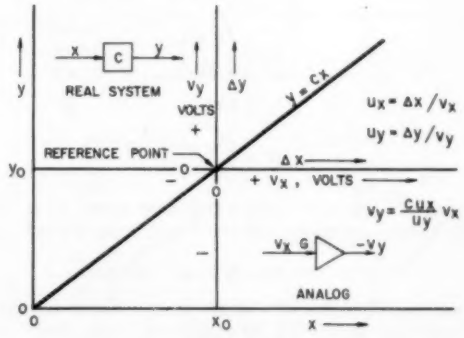


FIG. 1 FIXED COEFFICIENT

In a typical linear analog representation, a steady-state operating point of the simulated system is selected, and the variables in the system are mutually related by the fixed coefficients and dynamic relations that exist at that point. The steady-state point is normally zero volts for all variables. Excursions of the variables in response to applied disturbances occur as positive or negative delta amounts representing variations above or below the defined steady-state point. Take, for example, two variables x and y which are related by a coefficient C. If the initial steady-state points for these variables are constant terms defined as x_0 and y_0, then

$$x = x_0 + \Delta x \quad [1]$$

$$y = y_0 + \Delta y \quad [2]$$

where Δx and Δy are the excursions about the steady-state point. Relating the variables by the coefficient C, the following relations exist

$$y = C\,x \quad [3]$$

$$y_0 = C\,x_0 \quad [4]$$

$$y_0 + \Delta y = C\,(x_0 + \Delta x) \quad [5]$$

$$\Delta y = C\,\Delta x \quad [6]$$

Various methods may be used to apply scale factors which will express the variables in terms of analog voltages. A typical method uses scale units which define the amount of each variable per volt on the analog. The units are defined as constant terms u_x and u_y which relate the delta excursions to analog voltages according to the expressions

$$u_x = \frac{\Delta x}{v_x} \quad [7]$$

[1] Project Engineer, Curtiss-Wright Corporation, Propeller Division.

Contributed by the Instruments and Regulators Division of THE AMERICAN SOCIETY OF MECHANICAL ENGINEERS and presented at the ASME-AIEE Conference on Nonlinear Control Systems, Princeton, N. J., March 26–28, 1956.

NOTE: Statements and opinions advanced in papers are to be understood as individual expressions of their authors and not those of the Society. Manuscript received at ASME Headquarters, January 19, 1956. Paper No. 56—IRD-11.

$$u_y = \frac{\Delta y}{v_y} \quad [8]$$

In these equations v_x and v_y are variable voltages representing Δx and Δy on the analog equipment. The value of each scale unit is determined by the amount of delta excursion which is to be represented by the maximum analog voltage. The dimensions of the scale units will be determined by the dimensions of the corresponding variables. For example, if variable x has dimensions of rpm, then scale unit u_x has dimensions of rpm/volt.

In using the scale units to calculate the gain settings for analog amplifiers, the significant relations from Equations [6], [7], and [8] are

$$u_y v_y = C u_x v_x \quad [9]$$

$$v_y = \left(\frac{C u_x}{u_y}\right) v_x \quad [10]$$

$$v_y = G v_x \quad [11]$$

$$G = \frac{C u_x}{u_y} \quad [12]$$

The term G is the gain value which must be set on an analog amplifier to represent the coefficient C with scale factors included.

Function of a Single Variable

A nonlinear function of a single input variable represents a coefficient that varies with the magnitude of the input. Fig. 2 illustrates such a function. Several methods are available for matching this type of curve on analog equipment. These include

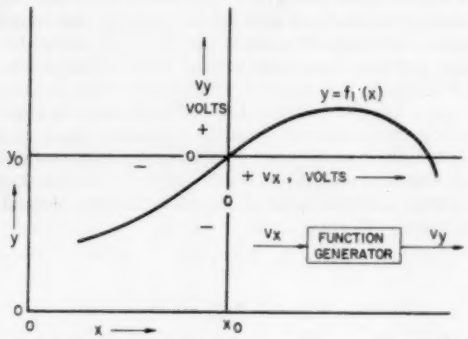


Fig. 2 Function of One Variable

servo-operated devices to move pickup resistors across wire which has been cemented to a graph, photoformers using a mask on the face of a cathode-ray tube, function-fitter components using diode circuits to make curves consisting of several straight-line segments, and polynomial equations to express the curve mathematically. The principles which are used here apply to any of these methods but are particularly intended for the function fitters which cover the full range of both positive and negative analog voltages. By using this full range of voltage to cover the expected excursion of each variable, the best accuracy is obtained from the analog equipment.

To apply a curve of this type to the analog, a reference point is selected on the curve at the approximate center of the expected range of variation. The values of the variables at this point are the reference values x_0 and y_0. Values for the scale units are determined in the same manner as the fixed coefficient, by the amount of excursion which is to be represented for each variable divided by the maximum analog voltage. Once the scale units have been selected, voltage axes may be applied to the plotted curve with their origin at the reference point. The voltage values are determined from the relations

$$v_x = \frac{x - x_0}{u_x} \quad [13]$$

$$v_y = \frac{y - y_0}{u_y} \quad [14]$$

Once the voltage values have been applied to the curve, scaling has been completed for the analog. When the function-fitting component is set up to represent the plotted curve in terms of volts, the correct system relation will be obtained in the analog circuit.

Translation of Curves

Plots of nonlinear functions of two input variables generally show a similarity or regular progression among the lines in the family of curves. In some cases this appears to be a translation or shifting of the curves, while in others a rotation, or fanning out of the curves is apparent. Procedures for accomplishing such translations and rotations on analog equipment make possible the matching of families of curves in a simple graphical manner.

The most frequently encountered case is that of translation. The analog mechanism for causing such shifts is to add voltages to the input and output of the function-fitting component which generates the basic shape of the curve.

Vertical Shifting. The effect of adding a voltage to the output of a function is illustrated by the plot in Fig. 3. The obvious result is that the curve is displaced vertically by an amount equal to the added voltage.

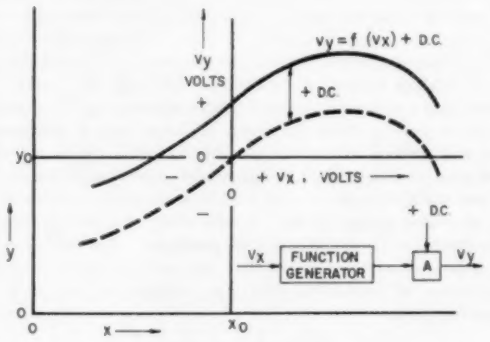


Fig. 3 Vertical Shifting

Horizontal Shifting. The effect of adding a voltage to the input of a function is illustrated in Fig. 4. In this case the curve is displaced horizontally, but an amount opposite in sign to the added voltage. A positive voltage added to the input causes the curve to shift to the left, or negative direction.

Combining Horizontal and Vertical Shifts. Since a mechanism is available for causing translation of curves in any desired direction, it is only necessary to control the amount of translation in response to a second input variable in order to match families of curves which are nonlinear functions of two input variables.

Fig. 5 illustrates the circuit required to produce combined horizontal and vertical translation in response to a second input variable. Two arrangements are shown: A block diagram, and a typical analog representation. In the latter the change of sign

produced by each analog amplifier is included. This circuit develops a voltage v_z which is related to one input v_x by the function

$$v_z = f_1(v_x) \quad [15]$$

and is translated horizontally and vertically by the second variable input voltage v_y. The horizontal shifting function is f_2 and the vertical shifting function is f_3. Notice that in the analog diagram, the horizontal function need not be expressed in the negative sense since the reversal in sign of the analog amplifier causes the necessary negative action.

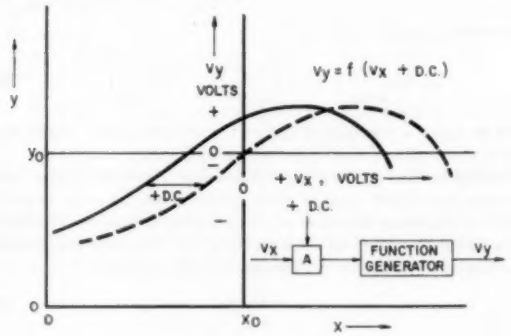


Fig. 4 Horizontal Shifting

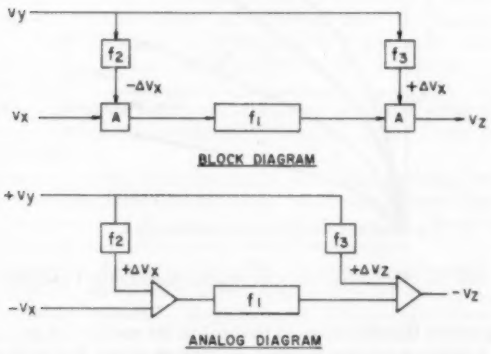


Fig. 5 Combined Horizontal and Vertical Translation

Determination of Shifting Functions. The method of determining the horizontal and vertical shifting functions is the most important step, and greatly influences the degree of accuracy which is obtained. To illustrate the procedure, the plot shown in Fig. 6 provides a good example of a three-dimensional function such as may be found in turbojet-compressor relations. As with the function of a single variable, a reference point is selected in the approximate center of the family of curves. All variables have zero voltage at this point. It is desirable, but not essential, that this point fall on one of the plotted curves. This line then represents zero volts for v_y, the second variable.

A transparent overlay having a single basic curve is plotted so that by successive shifts, it can be made to match each curve of the family. As the curves become steeper on the right side of the plot, they match the lower portion of the basic curve. In the less steep region of the left side, they match the upper portion of the basic curve. In matching the curves, care must be taken to keep the grid lines of the overlay parallel with the grid lines of the function being matched.

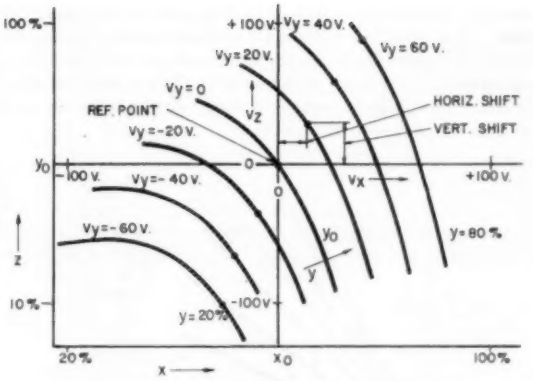


Fig. 6 Typical Three-Dimensional Function

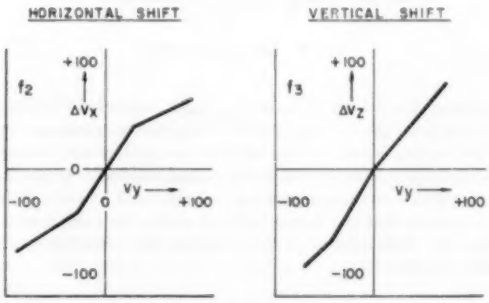


Fig. 7 Shifting Functions

To determine the shifting functions, the reference point is marked on the basic curve of the overlay. As this curve is shifted to each line of the family of curves, the point to which the reference point has moved is marked on each line. These points define the required amount of vertical and horizontal shift. As shown in Fig. 7, the amount of horizontal shift may be plotted as a variation of input voltage v_x for each increment of v_y to form f_2; and similarly the vertical shift f_3 is plotted as the variation of v_z corresponding to each increment of v_y. The shifting functions generally are not elaborate curves, and frequently can be matched by lines having two or three straight-line segments which require only simple limiter or diode circuits. Undesired nonlinearities obtained in first plots of the shifting functions can often be eliminated without undue sacrifice of accuracy by a second matching of the overlay to the curves, with judicious weighting in the desired direction where necessary.

When the basic curve of the overlay is applied to the analog circuit as f_1, and horizontal and vertical shifting curves are applied as f_2 and f_3, the desired three-dimensional relation will be generated. Interpolation between lines of the family of curves will follow the shapes of the shifting functions.

Rotated Curves

Families of curves which show a definite angular change between lines, or fan out from the origin, may best be represented by including rotation in the analog representation. This requires the use of multiplier components in one of several manners.

Direct Multiplication. In order to describe the analog methods for simulating curves which include multiplication, it is first necessary to establish the procedure for handling multiplication

when the delta and reference method is used for expressing variables. A graphical representation of multiplication is shown in Fig. 8. This plot illustrates the first quadrant of the product xy and consists of a series of straight lines radiating from the origin. Reference lines x_0 and z_0 also are plotted to indicate the zero-voltage axes of the analog simulation.

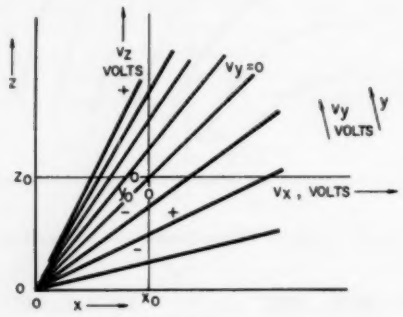


FIG. 8 MULTIPLICATION

Although the reference, or zero-voltage, point of the problem is displaced from the real origin when using the reference and delta method, multiplication of one variable by another can be accomplished easily. The process which is used improves the accuracy, since a portion of the computation is performed by normal linear equipment so that any inaccuracies in multiplier components are minimized. The method is illustrated by the following example:

Real equation

$$z = x\,y \qquad [16]$$

In reference form

$$(z_0 + \Delta z) = (x_0 + \Delta x)\,(y_0 + \Delta y) \qquad [17]$$

This becomes

$$z_0 + \Delta z = x_0 y_0 + x_0 \Delta y + y_0 \Delta x + \Delta x \Delta y \qquad [18]$$

Eliminating the constant values

$$z_0 = x_0 y_0 \qquad [19]$$

results in the expression

$$\Delta z = x_0 \Delta y + y_0 \Delta x + \Delta x \Delta y \qquad [20]$$

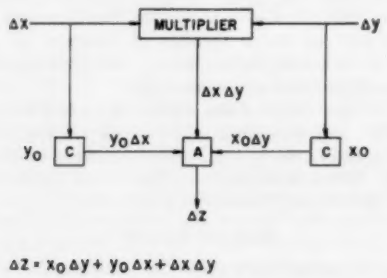


FIG. 9 MULTIPLICATION

The analog circuit required for multiplication is shown in the block diagram of Fig. 9.

In analog form the relations become

$$u_z v_z = x_0 u_y v_y + y_0 u_x v_x + (u_x v_x)(u_y v_y) \qquad [21]$$

$$v_z = \left(\frac{x_0 u_y}{u_z}\right) v_y + \left(\frac{y_0 u_x}{u_z}\right) v_x + \left(\frac{u_x u_y}{u_z}\right) v_x v_y \qquad [22]$$

The voltage v_z is thus the summation of three terms. The first two require only linear coefficients, while the third requires a multiplication as well as a coefficient.

Division. By slight revisions in the arrangement of terms, division also can be accomplished in a very simple manner.

For a real equation of the form

$$z = x/y \qquad [23]$$

the analog equation becomes

$$v_z = \left(\frac{u_x}{u_z y_0}\right) v_x - \left(\frac{u_y z_0}{u_z y_0}\right) v_y - \left(\frac{u_y}{y_0}\right) v_y v_z \qquad [24]$$

This is again a summation of two terms using linear coefficients and a third using a coefficient and a multiplier.

Multiplication of a Single Nonlinear Function. The most obvious and direct type of three-dimensional function is one in which a nonlinear function of one variable is directly multiplied by a second variable as shown in Fig. 10. A three-dimensional function of this nature is described by the equation

$$z = Cy\,f(x) \qquad [25]$$

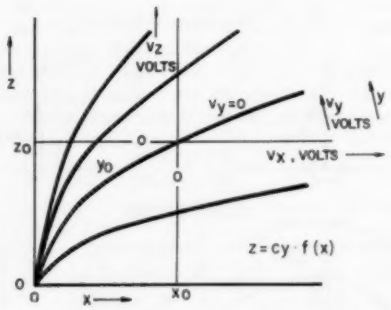


FIG. 10 MULTIPLICATION OF SINGLE NONLINEAR VARIABLE

To perform this operation on the analog, the method for developing a single nonlinear function is applied to the first variable. The output of this computation becomes one of the inputs to the multiplying arrangement which has been described. The second variable is applied to the other multiplier input through a suitable coefficient to produce the necessary angular variation between lines in the family of curves.

General Method for Rotated Functions. Many functions that involve rotation can be matched by generating a series of rotated straight lines by means of multiplication, and adding a nonlinearity to them. An example of such a function is illustrated in Fig. 11. To determine the analog relations for matching curves of this type, an overlay method is again used. By shifting and rotating the overlay, a series of radiating lines is established which defines the necessary multiplication. The basic nonlinear function, and the necessary shift and rotation functions are determined by the relation of these radiating lines to the desired family of curves. The following steps establish the multiplication and matching functions in voltage terms for application to analog equipment.

Basic Nonlinear Function, f_1. In a three-dimensional function, $z = f(x, y)$, establish a reference point on one basic line in the approximate center of the family of curves, and assign voltage

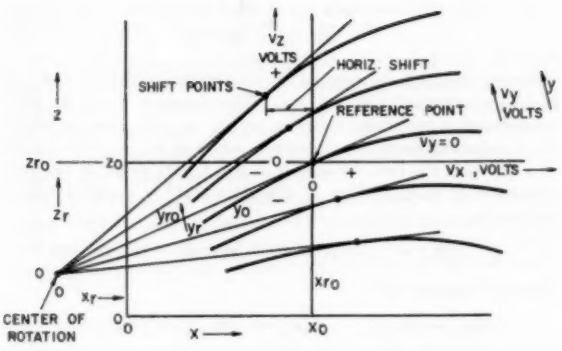


FIG. 11 ROTATED THREE-DIMENSIONAL FUNCTION

scales to each variable. All voltages are zero at the reference point.

Draw a tangent to the basic line at the reference point. The difference between the basic line and the tangent in terms of voltage v_z at each value of v_x defines the basic nonlinear function, f_1

$$v_{z1} = f_1 (v_x) \quad [26]$$

Center of Rotation and Shift Points. Copy the basic line with the reference point and the tangent line on a transparent overlay.

Shift and rotate the overlay to match each line of the family of curves. Establish a point of intersection for the tangents which will satisfy all curves. This is the "center of rotation" for the multiplication.

On each line of the family of curves, mark the point where the reference point of the overlay falls when matched. These are "shift points."

Horizontal Shifting Function, f_2. A horizontal shifting function is established from the horizontal magnitude of each "shift point" expressed as a change of v_x for each value of v_y. As in the case of translation, this function is negative in sign.

Multiplication Equation. Establish, in terms of the variables x and z, shifted scales which have the same units but originate at the "center of rotation." Magnitudes on these scales are designated x_r and z_r.

Each "shift point" will now have co-ordinates in terms of x_r and z_r. Those for the reference point will be x_{r0} and z_{r0}.

Solve for y_r, the value of the tangent at each shift point from the expression

$$y_r = \frac{z_r}{x_r} \quad [27]$$

at the reference point, the tangent will be

$$y_{r0} = \frac{z_{r0}}{x_{r0}} \quad [28]$$

Rotation Function, f_3. Multiplication about the "center of rotation" is now defined. Since this is not necessarily proportional to the second input voltage v_y, a rotation voltage v_r is established. This is related to v_y by a function f_3 in a manner similar to the vertical shifting function used in translation, and is found as follows:

Establish a scale unit u_r for the rotation voltage. This unit is based on the maximum excursion of the voltage v_y and the value of the tangent y_r at that condition, using the expression

$$u_r = \frac{y_{r\,max} - y_{r0}}{v_{y\,max}} \quad [29]$$

Determine the voltage v_r corresponding to the tangent y_r for each line of the family of curves from the expression

$$v_r = \frac{y_r - y_{r0}}{u_r} \quad [30]$$

Plot v_r versus v_y to obtain the rotation function f_3.

Analog Arrangement for Rotation. Fig. 12 illustrates the analog circuit required to represent a function of two input variables

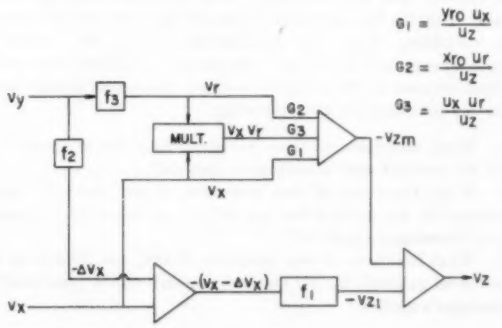


FIG. 12 ANALOG ARRANGEMENT FOR ROTATED FUNCTIONS OF TWO INPUT VARIABLES

involving rotation. The multiplication is in terms of one input voltage v_x and the rotation voltage v_r in accordance with the equation

$$v_{zm} = \left(\frac{y_{r0}u_x}{u_z}\right) v_x + \left(\frac{x_{r0}u_r}{u_z}\right) v_r + \left(\frac{u_x u_r}{u_z}\right) v_x v_r \quad [31]$$

The total output voltage v_z is the summation of the voltages developed by function f_1 and the multiplication

$$v_z = v_{z1} + v_{zm} \quad [32]$$

The additional shifting and rotation functions f_2 and f_3 are included at the appropriate points to add to v_z and develop v_r in response to the second input voltage v_y.

APPLICATIONS

By using the principles of translation or rotation which have been developed here, some combination can be obtained to match almost any nonlinear function of two input variables which has reasonably uniform progression among the lines. Using these methods, it has been possible to match turbojet-compressor curves within approximately 1/2 per cent of full scale within the working range, with the error increasing to the order of 2 per cent at the limits of the function. The accuracy of the simulation can be checked graphically before applying the circuits to the analog equipment.

Since only standard analog equipment is used, the circuits may be drawn in a normal manner once the constants have been determined. Interconnections among analog components can be made easily from the circuit diagrams by anyone familiar with analog connections, since no special revisions to components are required. By using all-electronic function generators and multipliers, no objectionable dynamics are added to the problem.

Extension of these methods to functions of three input variables is also possible. If such a function can be plotted as a series of families of curves, the transitions between families can be determined. Suitable shifts can then be applied in much the

same manner as used to convert functions of a single variable to functions of two. Obviously the amount of equipment required and the complexity of the problem increase when an additional input variable is introduced.

Discussion

E. S. SHERRARD.[2] Other methods than those described in the paper may be used for generating a function of two variables with electronic components. Methods for handling quite arbitrary functions of two variables have been described by Jerrard and Jacobi,[3] and by Philbrick.[4] The isocline method of Meissinger[5] is useful for generating a particular class of functions of two variables. Since the equipment used by the author is simple, cheap, and much more frequently available than is the equipment used in Meissinger's method, the writer would appreciate his comment on the following:

1 What functions of two variables may be generated by both his method and Meissinger's method?

2 What functions of two variables, if any, may be easily generated by his method but are difficult or impossible to generate by Meissinger's method?

3 What functions of two variables, if any, are difficult or impossible to generate by his method, but are easily generated by Meissinger's method?

The author's method is capable of generating exactly or approximately the family of curves $z(x, y_i)$, according to whether Equations [33] and [34] or Equations [35] and [36] are exactly or approximately satisfied

$$z(x, y_i) = f_1(w) + f_2(\delta y_i) \text{ (see Fig. 5)} \quad [33]$$

$$w = x + f_2(\delta y_i), \quad \delta y_i = y_i - y_0 \quad [34]$$

$$z(x, y_i) = f_1(w) + x f_2(\delta y_i) \text{ (see Fig. 11)} \quad [35]$$

$$w = x + f_2(\delta y_i), \delta y_i = y_i - y_0. \quad [36]$$

where the "reference function" $f_1(w)$ is taken equal to $z(x, y_0)$, y_0 being the reference value of y, the parameter of the family. In the foregoing equations $f_2(\delta y_i)$ is the horizontal shift function and $f_2(\delta y_i)$ is the vertical shift function illustrated in Fig. 7 of the paper. The author's method of determining these shift functions is a cut-and-try procedure which requires the drawing of an overlay of $f_i(w)$. Such a method may be the best approach if Equations [33] and [34] or [35] and [36] are satisfied approximately rather than exactly. If these equations are not satisfied approximately, the cut-and-try procedure may be both tedious and ineffective.

A simple test for the satisfaction of Equations [33] and [34] may be made by using the first differences with respect to x of $z(x, y_i)$ and the first differences of $f_1(w)$ with respect to w. If Equation [33] is satisfied, then differentiation of Equation [33] yields

$$\frac{\partial z(x, y_i)}{\partial x} = \frac{df_1(w)}{dw} \quad [37]$$

Equation [37] will be satisfied for corresponding values of w and x, (w_1, x_1), (w_2, x_2), (w_3, x_3), etc. The values of x_1, x_2, x_3, etc., corresponding to chosen values of w_1, w_2, w_3, etc., may be found by determining from the difference tables of $z(x, y_i)$ the values of x at which the first differences of $z(x, y_i)$ are equal to the first difference of $f_1(w)$. Then, if Equation [34] is satisfied $f_2(\delta y_i)$ is given by

$$f_2(\delta y_i) = w_1 - x_1 = w_2 - x_2 = w_3 - x_3, \text{ etc.}$$

Also, for Equation [33]

$$f_2(\delta y_i) = z(x_1, y_i) - f_1(w_1) = z(x_2, y_i) - f_1(w_2), \text{ etc.}$$

The constancy of $f_2(\delta y_i)$ and $f_2(\delta y_i)$ as w and x assume the values of (w_1, x_1), (w_2, x_2), (w_3, x_3), etc., determines whether or not Equations [33] and [34] are good or poor approximations for the given function $z(x, y_i)$.

If $z(x, y_i)$ is satisfied by Equations [35] and [36], the second differences of $z(x, y_i)$ and $f_1(w)$ may be used to test Equations [35] and [36]. Differentiation of Equation [35] yields

$$\frac{\partial^2 z(x, y_i)}{\partial x^2} = \frac{d^2 f_1(w)}{dw^2} \quad [38]$$

From the tables of second-order differences, corresponding values (w_1, x_1), (w_2, x_2), (w_3, x_3), etc., that satisfy Equation [38] may be determined. Then, $f_2(\delta y_i)$ is again equal to

$$w_1 - x_1 = w_2 - x_2 = w_3 - x_3, \text{ etc.}$$

and $f_2(\delta y_i)$ is determined from Equation [35] with (w, x) set equal to (w_1, x_1), (w_2, x_2), (w_3, x_3), etc. As before, the constancy of $f_2(\delta y_i)$ and $f_2(\delta y_i)$ will determine whether Equations [35] and [36] are a good or a poor approximation of $z(x, y_i)$.

An accurate graphical solution of Equation [37] may be made by using the well-known "mirror method" of graphical differentiation. To do so, one may determine the slope of $f_1(w)$ at a chosen value of w, w_1. x_1 is the value of x at which the slope of $f(x, y_i)$ is equal to the slope of $f_1(w)$ at $w = w_1$. Other pairs of corresponding values (w_2, x_2), (w_3, x_3), etc., may be found by the same method. Then, as before,

$$f_2(\delta y_i) = w_1 - x_1, w_2 - x_2, w_3 - x_3, \text{ etc.}$$

and $f_2(\delta y_i)$ is found from Equation [33] for (w, x) equal to (w_1, x_1), (w_2, x_2), (w_3, x_3), etc. There does not seem to be any simple and accurate graphical method for determining corresponding values of w and x that satisfy Equation [38].

AUTHOR'S CLOSURE

The comments of Mr. Sherrard are greatly appreciated. The methods which he describes add a considerable improvement in mathematical treatment to the graphical principles which have been described.

In answer to his specific questions concerning the differences between the Meissinger method and that described here, the following comparisons may be made:

1 Both methods can generate functions which have constant spacing between isoclines.

2 Functions which contain intersecting lines of y = const may be generated easily by the method shown in the paper. This is a more difficult case for the Meissinger method than that of nonintersecting curves having a laminated pattern.

3 The method described here cannot generate functions in which the curves of y = const are greatly dissimilar in basic

---

[2] Analog Systems Section, Data Processing Systems Division, U. S. Department of Commerce, National Bureau of Standards, Washington, D. C.

[3] "Generation of a Function of Two Variables," by R. P. Jerrard and G. T. Jacobi, presented at the Annual Conference of the Association for Computing Machinery, Cambridge, Mass., September, 1953.

[4] "Continuous Electric Representation of Non-Linear Functions of n-Variables," by G. A. Philbrick, A Palimpsest on the Electronic Analog Art, G. A. Philbrick Researches, Inc., 1955.

[5] "An Electronic Circuit for the Generation of Functions of Several Variables," by H. F. Meissinger, Institute of Radio Engineers Spring Meeting, 1955.

shape. The Meissinger method can generate some cases of this nature using a three-channel diode network.

The basic Meissinger cases of convergent radiating isoclines develop families of curves which expand outward from the convergent point. These cannot be generated directly by the methods which have been described here. However, by substitution of multiplications for the summations shown in Fig. 5, functions of this nature can be represented.

The other methods for generating functions of two variables which Mr. Sherrard has mentioned are certainly more general and will provide more uniform accuracy than that described here. As he points out, "map readers" of this nature are also much more expensive, and involve considerably more complexity of equipment.

We have also used a map-reading technique of this nature for simulating difficult functions. The method generates a family of arbitrary curves, but requires a function generator for each curve, a multiplier to interpolate between each pair of lines, and three amplifiers per curve. Although this is effective for an occasional three-dimensional function, such a technique becomes prohibitive if many functions must be simulated in a given problem. The method described in the paper has been presented as a useful procedure where simplicity and economy of equipment are desired.

Basic Methods for Nonlinear Control-System Analysis

By T. M. STOUT,[1] LOS ANGELES, CALIF.

The important phase-plane and describing-function methods are explained with the help of several simple examples. It is shown that the phase-plane method is best adapted to transient analysis of second-order systems and that the describing-function method is intended primarily for stability analysis of higher-order systems. Details and extensions of both methods are discussed in two appendixes, and a number of less important methods are mentioned.

NUMEROUS methods which could be used to analyze nonlinear control systems are described in the literature on nonlinear mechanics and numerical analysis (see, for example, references 1–7);[2] not all of these methods, however, have been applied to control problems. In addition, control engineers have invented new methods or modified existing methods to fit their needs. The control-system analyst should be familiar with as many methods as possible, since the various methods cannot be used interchangeably and some are better suited to particular kinds of problems than others.

Some methods are inherently approximation methods, while others are capable of arbitrary refinement. Some methods work best (or at all) for systems which are only slightly nonlinear, while others can accommodate any degree of nonlinearity. Some methods give the response for a particular input or set of initial conditions, others indicate whether a system is stable or not, others give the amplitude and frequency of a sustained oscillation if one occurs, and still others give a birds-eye view of over-all system behavior. Some methods are adaptations or extensions of methods used for linear systems, while others are completely different.

In a short paper, we cannot list and explain all available methods. The paper therefore emphasizes the two methods which are most widely used in control-system analysis, the phase-plane method and the describing-function method. By means of some simple examples, we will attempt to show how the methods are used, what sort of information they give, and their advantages and limitations. Additional information concerning these two methods is given in two appendixes. Several less important methods also will be mentioned.

Phase-Plane Method

Phase-plane analysis of dynamic systems goes back at least 50 years to work of physicists and astronomers. Extension of these techniques to control-system problems is, of course, more recent. Some of the earliest applications can be found in work of Doll (8), Hurewicz (9), MacColl (10), and Weiss (11).

The phase plane, whose co-ordinates are usually the system error and error rate, provides both a means for representing phenomena in nonlinear systems and a means for detailed analysis of such systems. If the system is described by a second-order differential equation and if the input is a step or ramp function, the state of the system at any time is completely specified by a point in the phase plane. A succession of such points, joined by a continuous curve to show the history of the system, is called a "trajectory."

In the examples which follow, we assume that the trajectories are known, either by experiment or calculation. We adopt this approach to focus attention on use of the phase plane and the information that can be obtained from it. The real value of the phase-plane method lies, of course, in the fact that trajectories showing system behavior can be constructed even when the variables cannot be found directly as functions of time. A number of methods for constructing trajectories are explained in Appendix 1.

Examples

The following examples all concern a positioning servomechanism with an output member characterized by inertia (J) and viscous friction (f). The torque used to position the output member is, in general, a nonlinear function of the error between the desired and actual positions of the output shaft.

1 *Linear System.* Any method for analyzing nonlinear systems also must work for linear systems, although the converse is unhappily not true. We begin by considering a system in which the torque is proportional to the error. The torque-error relation is expressed in this case by the single equation

$$T = Ke \quad [1]$$

The system trajectories have the possible forms shown in Fig. 1. We suppose that the system has been subjected to a step input and starts at point A. If the system damping (proportional to f) is small, the error reaches zero in an oscillatory manner, shown by the spiral trajectories of Fig. 1(A). If the system damping is large, the error decreases to zero without overshoot, as shown in Fig. 1(B). If the system damping is zero, continuous oscillations occur; the system returns to its starting point on a closed elliptical trajectory, as shown in Fig. 1(C).

Since a positive error rate always corresponds to an increasing error and a negative error rate to a decreasing error, the arrows along the trajectories always point to the right in the upper half of the plane and to the left in the lower half.

This system is stable for all parameter values (except $f \leqslant 0$), since all trajectories end at the origin, and the steady-state error for step inputs is zero. Because the system is linear, trajectories for other step magnitudes can be drawn by a simple scaling process, expanding or contracting the given trajectories to pass through the specified initial point.

2 *Saturating Linear System.* The system considered in the first example becomes nonlinear if the torque becomes constant ($\pm T_m$) whenever the absolute error exceeds a certain value (e_s),

---

[1] Ramo-Wooldridge Corporation; this paper was written while the author was employed by the Schlumberger Instrument Company, Ridgefield, Conn.

[2] Numbers in parentheses refer to the Bibliography at the end of the paper.

Contributed by the Instruments and Regulators Division of The American Society of Mechanical Engineers and presented at the ASME-AIEE Conference on Nonlinear Control Systems, Princeton, N. J., March 26–28, 1956.

Note: Statements and opinions advanced in papers are to be understood as individual expressions of their authors and not those of the Society. Manuscript received at ASME Headquarters, January 6, 1956. Paper No. 56—IRD-9.

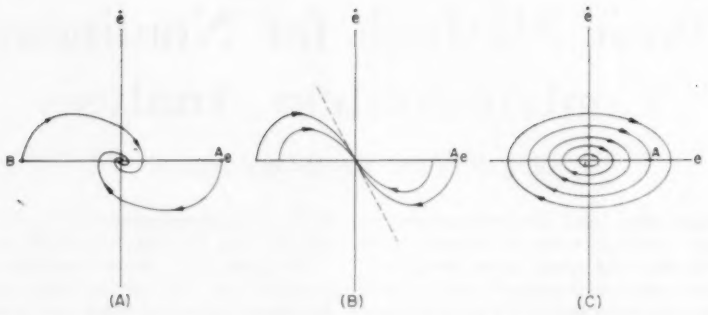


FIG. 1 PHASE-PLANE REPRESENTATION OF LINEAR-SYSTEM BEHAVIOR: (A) SMALL DAMPING; (B) LARGE DAMPING; (C) NO DAMPING

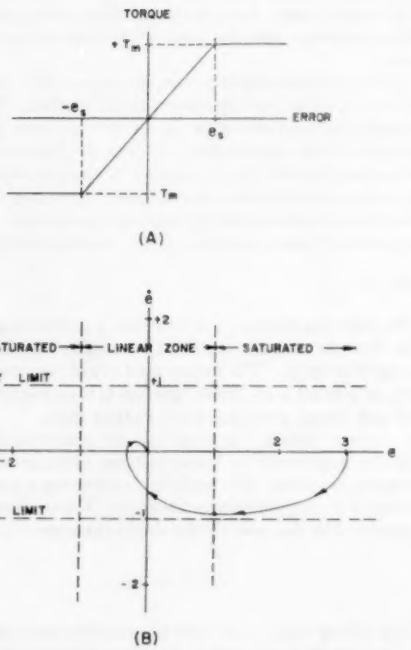


FIG. 2(A) TORQUE-ERROR RELATION IN SATURATING LINEAR SYSTEM; (B) PHASE-PLANE REPRESENTATION OF STEP RESPONSE

as indicated in Fig. 2(A). The torque-error relation is now given by three equations

$$T = \begin{cases} +T_m & e > e_s \\ K\,e & -e_s < e < e_s \\ -T_m & e < -e_s \end{cases} \quad \ldots\ldots\ldots\ldots [2]$$

The phase plane likewise may be divided into three regions, a linear region bounded by the lines $e = \pm e_s$ and two saturated regions. In the linear region, the trajectories have the possible shapes of Fig. 1, the exact shape depending on the parameters of the system. In the saturated regions, the trajectories approach a limiting error rate, $\dot{e} = \pm T_m/f$.

A typical trajectory, obtained by fitting together trajectories appropriate to the respective regions, is shown in Fig. 2(B) for the case $T_m = J = f = e_s = 1$. The system is again stable and all trajectories will end at the origin. Since the trajectories cannot enter the linear region with an error rate greater than the limiting rate, the overshoot cannot be much greater than that shown, about one third of a unit for a three-unit step input.

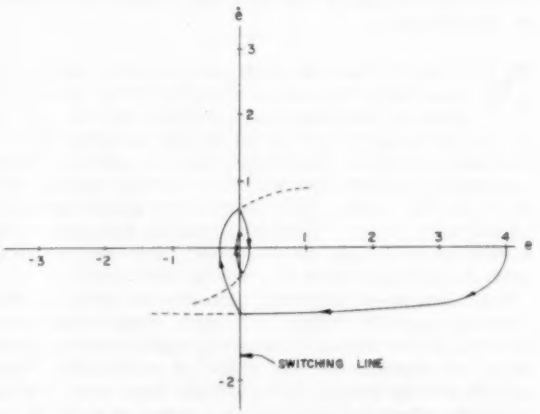


FIG. 3 PHASE-PLANE REPRESENTATION OF RELAY SYSTEM

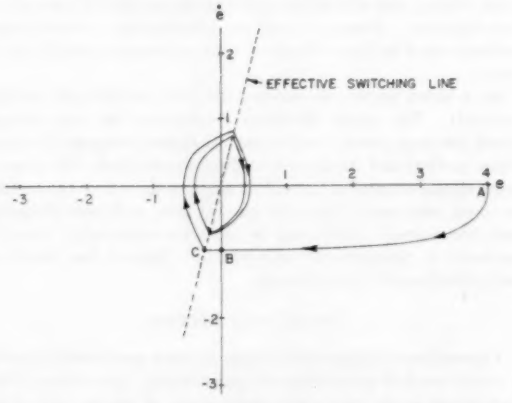


FIG. 4 PHASE-PLANE REPRESENTATION OF RELAY SYSTEM WITH TIME DELAY

3 *Relay System.* If e_s in the previous system is made smaller and smaller by increasing K, the system becomes effectively a relay system in which the torque-error relation is

$$T = \begin{cases} +T_m & e > 0 \\ -T_m & e < 0 \end{cases} \quad \ldots\ldots\ldots\ldots [3]$$

The phase plane is now divided into two parts with relay operations occurring on the dividing line between the two regions, $e = 0$. The trajectory following a four-unit step is shown in Fig. 3. The system is again stable and the maximum possible overshoot is about one-third unit.

4 *Relay System With Time Delay.* We suppose as before that the relay starts to operate when the error is zero, but that torque reversal actually occurs T_d time units later. Since the distance traveled in a fixed time is proportional to the velocity, the effective switching line is

$$e - T_d\dot{e} = 0 \quad \ldots\ldots [4]$$

As shown in Fig. 4, switching is initiated at point B but torque reversal does not occur until the trajectory reaches point C. The system is now unstable with all trajectories ending in a closed curve or "limit cycle." The amplitude of the limit cycle in this case is about one-third unit; the period of the oscillation is not available directly but turns out to be about 3.7 time units.

5 *Relay System With Dead Zone.* In this case, we suppose that a finite error signal is needed to operate the relay but that relay operation is instantaneous. The torque-error relation is therefore

$$T = \begin{cases} +T_m & e > e_d \\ 0 & -e_d < e < e_d \\ -T_m & e < -e_d \end{cases} \quad \ldots\ldots [5]$$

The phase plane is again divided into three regions. In the regions corresponding to $+T_m$ and $-T_m$, the trajectories are similar to those of the saturating linear system or relay system (Examples 2 and 3). In the center zone where no torque is applied to the output member, the error rate decreases because of friction losses; the trajectories in this region are straight lines with a slope $-f/J$.

A typical trajectory is plotted in Fig. 5 for the case $e_d = 0.25$; the overshoot is about 0.4 unit and the final error about 0.125 unit. This system is stable in the sense that the error rate eventually reaches zero; there is, however, a steady-state position error less than or equal to e_d.

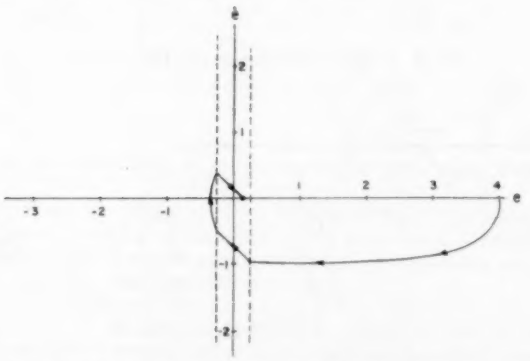


Fig. 5 Phase-Plane Representation of Relay System With Dead Zone

Comments

The phase-plane method is suitable for transient analysis of systems described by a single nonlinear differential equation of second order or by a succession of linear or nonlinear differential equations, each applicable in certain ranges of the variables. Several nonlinearities can be considered simultaneously. A phase-plane analysis can be made as accurate as desired and shows the general nature of the system behavior. In addition, the phase-plane plots give a direct indication of the degree of damping or amount of overshoot and the amplitude of any sustained oscillations which may occur.

The times associated with each point on a trajectory, needed to find the response time for step inputs or the period of any sustained oscillations, are not directly available; they can, however, be computed from the relation

$$t_{AB} = \int_{e_A}^{e_B} \frac{1}{\dot{e}}\, de \quad \ldots\ldots [6]$$

or by a graphical construction due to Diprose (13).

A unique relation between the system trajectories and the phase-plane co-ordinates exists only for second-order systems subjected to step or ramp inputs. The behavior of a second-order system with more general inputs can be computed by generalizations of the basic phase-plane techniques and recorded in the error-error rate plane. For higher-order systems, a phase space is required; projections of phase-space trajectories into the error-error rate plane may still be useful (12, 14, 36).

Describing-Function Method

The describing-function method rests on the work of Fourier and parallels techniques used for some time in the field of nonlinear mechanics. As a method for analysis of nonlinear control systems, it seems to have been developed independently and almost simultaneously by Tustin in England (15), Goldfarb in Russia (16), Oppelt in Germany (17), Dutilh in France (18), and Kochenburger in the United States (19). Numerous extensions and applications of the basic idea have been made in the past few years.

The describing function for a nonlinear element is found by ap-

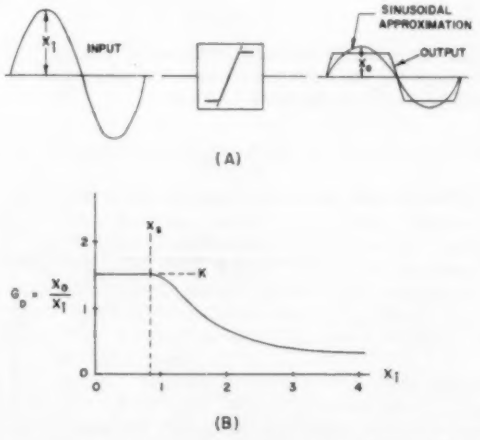


Fig. 6 Describing Function for Saturating Linear Element

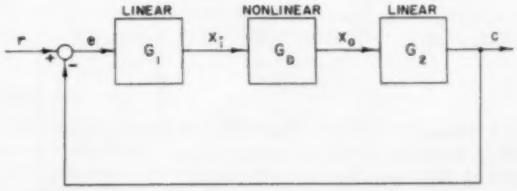


Fig. 7 Block Diagram of Control System With a Nonlinear Element

plying a sinusoidal input to the element and, using standard Fourier analysis techniques, determining the fundamental component of the distorted output. The describing function has two parts: (*a*) The output-input amplitude ratio, computed from the peak values of the fundamental output component and the sinusoidal input; and (*b*) the phase shift between the sinusoidal input and the fundamental output component.[3] For elements which are completely described by a curve relating the instantaneous input and output, the describing function depends on the magnitude of the input but is independent of its frequency.

The describing function for a saturating linear element is shown in Fig. 6. For inputs less than x_s, the describing function is K, the slope of the output-input curve. For larger inputs, the output magnitude is limited and the describing function decreases.

Use of describing functions is based on the block diagram shown in Fig. 7. For checking stability, we can assume the input (r) is zero. The condition for the existence of a sustained oscillation is

$$e = -c \quad [7]$$

which requires that

$$G_1 G_D G_2 = -1 \quad [8]$$

or

$$G G_D = -1 \quad [9]$$

if we let $G = G_1 G_2$. In these equations, G_1 and G_2 represent ordinary linear gain functions which are independent of amplitude but dependent on frequency, and G_D is the describing function for the nonlinear element. In our examples, G is given by

$$G(\omega) = \frac{1}{j\omega(Jj\omega + f)} \quad [10]$$

$$= \frac{1}{j\omega(j\omega + 1)} \quad [11]$$

and G_D represents the nonlinear torque-error relation.

To check for the existence of sustained oscillations, we will write Equation [9] in the equivalent form

$$G = -\frac{1}{G_D} \quad [12]$$

[3] For the examples used in this paper, the phase shift is zero.

We then plot G and $-1/G_D$ as separate curves in a standard polar plot showing the magnitude and phase of each quantity. Intersections of the two curves, if any exist, give the approximate magnitude and frequency of possible sustained oscillations. Further analysis is required to determine whether the intersections are points of convergent or divergent equilibrium (19).

Examples

1 *Linear System.* If there is no saturation, the torque-error relation is simply $G_D = K$. The curve $-1/G_D$ becomes the point $-1/K$. As shown in Fig. 8, $G(\omega)$ is infinite and has a phase angle of —90 deg at zero frequency, and approaches zero with a phase angle of —180 deg at infinite frequency. As is the case with the usual Nyquist criterion, the system is stable since the curve $G(\omega)$ does not enclose the critical point $-1/K$, regardless of the value of K.

2 *Saturating Linear System.* Fig. 6 indicates that saturation causes G_D to become less than K when the input to the saturating element is large. Thus $1/G_D$ is always equal to or greater than $1/K$, and $-1/G_D$ occupies the entire negative axis of the polar

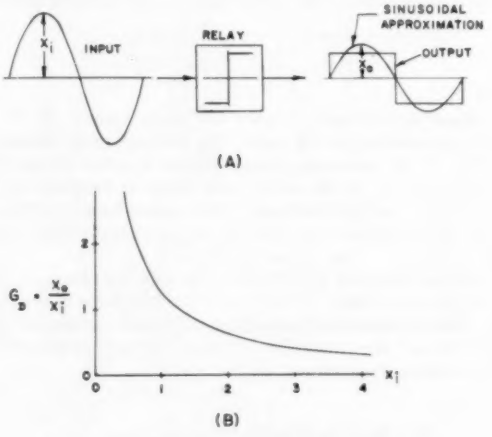


Fig. 9 Describing Function for Ideal Relay

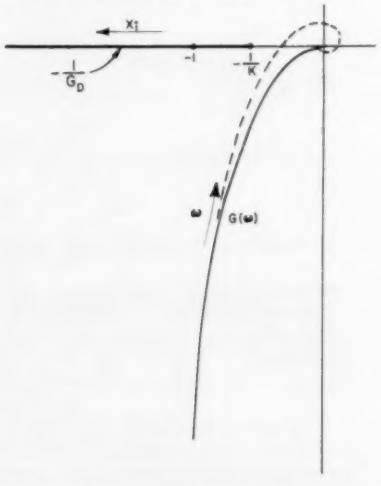


Fig. 8 Polar Plot for Linear System and Saturating Linear System

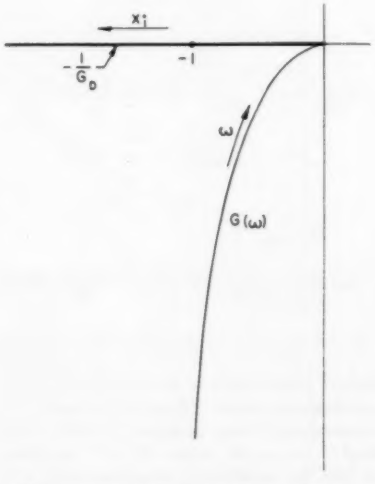


Fig. 10 Polar Plot for System With Ideal Relay

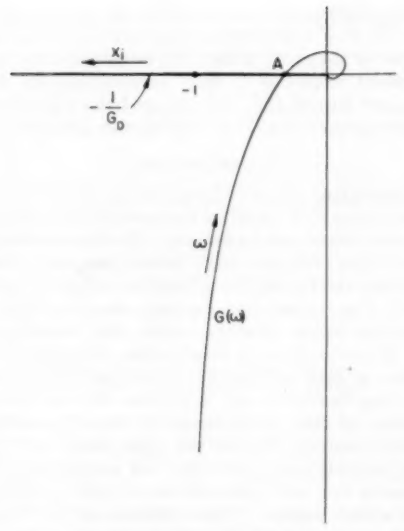


FIG. 11 POLAR PLOT FOR SYSTEM WITH IDEAL RELAY AND TIME DELAY

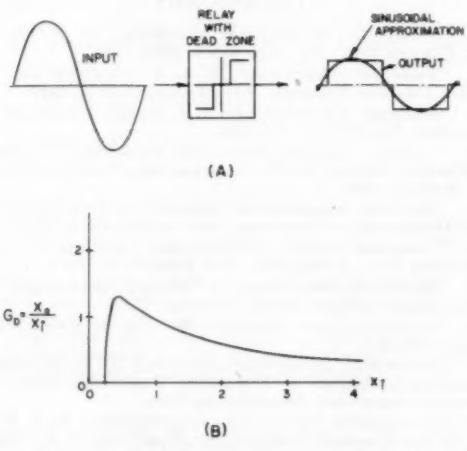


FIG. 12 DESCRIBING FUNCTION FOR RELAY WITH DEAD ZONE

plot to the left of the point $-1/K$, as shown in Fig. 8. The system is stable under the conditions shown. It could exhibit sustained oscillations if K were increased and if additional phase lags, indicated by the dashed line, were introduced.

3 *Relay System.* The calculation of the describing function for an ideal relay is based on Fig. 9(A). Since the relay is assumed to have no dead zone the relay output is a square wave for a sinusoidal input of even infinitesimal amplitude. Since the fundamental component of a square wave is $(4/\pi)$ (peak value), G_D is a constant divided by X_i, giving the hyperbola plotted in Fig. 9(B). Since G_D now ranges from zero to infinity, $-1/G_D$ occupies the entire negative axis of the polar plot, as shown in Fig. 10. For the $G(\omega)$ considered in these examples, the system is stable; with any slight additional phase shift, sustained oscillations would take place.

4 *Relay System With Time Delay.* Addition of time delay to the relay system can be represented by adding a phase shift proportional to frequency to $G(\omega)$, giving the curve shown in Fig. 11. The curves $G(\omega)$ and $-1/G_D$ intersect at point A, indicating the possibility of a sustained oscillation. In terms of the usual Nyquist criterion, all points on $-1/G_D$ between A and the origin are enclosed by the curve $G(\omega)$, so that the system is unstable in some sense for the small X_i amplitudes associated with G_D in this range. In similar fashion, the system is stable for points on $-1/G_D$ to the left of A which are not enclosed and which correspond to large X_i amplitudes. Point A is therefore a convergent equilibrium point, since larger amplitudes decay to A and smaller ones grow to A. Following any slight disturbance, this system will therefore oscillate with the approximate frequency and amplitude associated with point A.

5 *Relay System With Dead Zone.* If the relay has a dead zone, its output is obviously zero for small inputs. The describing function G_D is therefore zero for small inputs, increases rapidly for inputs slightly in excess of the dead zone, reaches a maximum, and then returns to zero, as shown in Fig. 12. The polar plot of Fig.

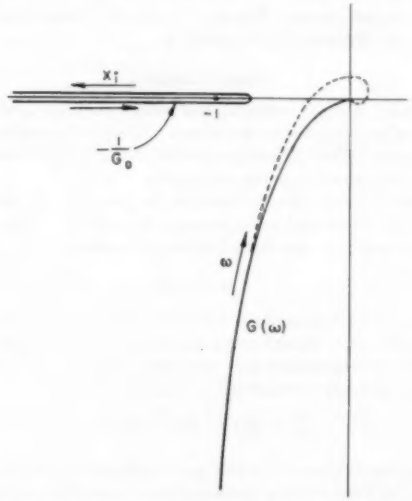


FIG. 13 POLAR PLOT FOR SYSTEM WITH RELAY HAVING DEAD ZONE

13 indicates that the system is stable and would, in fact, remain stable with the introduction of additional phase lags (indicated by the dashed line). With a smaller dead zone, however, the maximum value of G_D increases, $-1/G_D$ occupies a larger part of the negative axis and comes closer to the origin, thus creating the possibility of an intersection of $-1/G_D$ with the $G(\omega)$ curve shown by the dashed line.

In this particular case, any intersection of $-1/G_D$ with $G(\omega)$ is, in fact, two intersections. The intersection with the upper branch of $-1/G_D$ is a convergent equilibrium point (for reasons discussed in connection with Example 4) corresponding to a large amplitude oscillation. The other intersection, on the lower branch of $-1/G_D$, is a divergent equilibrium point and represents a sort of dividing line between disturbances which decay back inside the dead zone and those which build up into a stable oscillation.

Comments

The describing-function method gives a direct indication of the existence, amplitude, and frequency of possible sustained oscillations. In general, the errors in the predicted amplitude and frequency are not more than ten per cent (19). This accuracy is probably consistent with the original data and adequate for most

engineering purposes. On the other hand, the describing-function method is inherently an approximate method and may therefore fail occasionally to predict a sustained oscillation which actually does occur or may predict one which does not occur.

The proximity of the $-1/G_D$ and $G(\omega)$ curves gives a rough idea of the relative stability of the system and thus provides a very qualitative indication of the transient response to be expected. The curves also can be used to construct approximate frequency-response curves for the system; unlike the linear case, the output-input ratio is a function of amplitude as well as frequency.

In contrast with the phase-plane method, the describing-function method is applicable to systems in which the linear elements are described by differential equations of third or higher order. The effects of various compensation methods (which increase the order of the differential equations) can be examined by making appropriate modifications of the $G(\omega)$ curve, just as in the case of linear system design.

Some of the assumptions underlying the describing-function method are discussed in Appendix 2.

Other Methods

A number of other methods, less widely used, are available for analysis of nonlinear control systems. Most of the methods mentioned in the following paragraphs are useful for determining the response of systems to particular inputs.

Numerical integration techniques, in principle, can always be used if the differential equations can be written. Many of the numerical methods concern a first-order equation

$$\dot{x} = F(x, t) \quad [13]$$

or systems of such equations, although methods also exist for dealing directly with second-order equations. These methods are essentially extrapolation and correction schemes. The simplest procedure is linear extrapolation, using the relation

$$x(t + \Delta t) = x(t) + \Delta t \dot{x}(t) \quad [14]$$

The computed value of x can be used to obtain an average $\dot{x}$ in the interval Δt, from which a more accurate value of x can be found. Still better methods are described in texts on numerical analysis (5 to 7). These methods are somewhat tedious for hand use but are basic to automatic digital computers.

Equivalent methods in which the integrations are performed graphically (instead of numerically) have been proposed by Hsia (20) and Paynter (21).

A number of step-by-step methods combining numerical and graphical techniques have been proposed by Tustin (22), Madwed (23), and Stout (24). In these methods, the response of a linear element is computed from its input by numerical methods, making direct use of the superposition principle, and the nonlinearity is taken into account by a graphical construction.

Another adaptation of the superposition principle was used by Kahn (25) in a study of relay servomechanisms. Since the system following the relay was linear, its output could be found by superposition of step responses corresponding to the application and removal of a fixed input by the relay. The variables entering elements preceding the relay determine the times of relay operation but do not otherwise affect the system output.

In certain systems, such as the saturating or relay systems used in the examples, operation is governed by a number of linear differential equations applicable in restricted ranges of the system variables. This property, known as "piece-wise linearity," can be used to good advantage in phase-plane analysis and also can be used to construct response curves directly as functions of time. The response for any input is found by joining solutions of the linear equations at transitions from one mode of operation to another, the final conditions for one mode becoming the initial conditions for the next. Hazen (26) investigated the conditions for sustained oscillation in relay servomechanisms using this method, and Nichols (27) recently used the same approach to check the accuracy of a describing-function analysis.

Conclusions

The phase-plane and describing-function methods can furnish answers to many (but not all) of the performance questions which concern the control-system designer. The two methods complement each other, the phase-plane method being useful for second-order systems and the describing-function method for higher-order systems. The phase-plane method deals primarily with transient conditions, however, while the describing-function method is concerned with steady-state oscillatory conditions. Extensions of both methods have been and will be devised to remedy these limitations and to increase their usefulness.

A number of other methods are available for nonlinear control-system analysis. Unlike the phase-plane and describing-function methods these methods reveal general features of system behavior only after exhaustive investigation of all kinds and sizes of system inputs. These methods are therefore better adapted to analysis of specific systems than to system design or modification.

BIBLIOGRAPHY

1 "Introduction to Nonlinear Mechanics," by N. Minorsky, J. W. Edwards Bros., Inc., Ann Arbor, Mich., 1947.

2 "Theory of Oscillations," by A. A. Andronow and C. E. Chaikin, Princeton University Press, Princeton, N. J., 1949.

3 "Nonlinear Vibrations," by J. J. Stoker, Interscience Publishers, Inc., New York, N. Y., 1950.

4 "Ordinary Nonlinear Differential Equations in Engineering and Physical Sciences," by N. W. McLachlan, Clarendon Press, Oxford, England, 1950.

5 "Numerical Mathematical Analysis," by J. B. Scarborough, Johns Hopkins Press, Baltimore, Md., second edition, 1950.

6 "Numerical Solution of Differential Equations," by W. E. Milne, John Wiley & Sons, Inc., New York, N. Y., 1953.

7 "Numerische Behandlung von Differentialgleichungen," by L. Collatz, Julius Springer, Berlin, Germany, 1951.

8 "Automatic Control System for Vehicles," by H. G. Doll, U. S. Patent 2,463,362.

9 "Servos With Torque Saturation—Part II," by W. Hurewicz, Massachusetts Institute of Technology, Cambridge, Mass., Radio Laboratory Report 592, September 28, 1944.

10 "Fundamental Theory of Servomechanisms," by L. A. MacColl, D. Van Nostrand Company, Inc., New York, N. Y., 1945.

11 "Analysis of Relay Servomechanisms," by H. K. Weiss, *Journal of the Aeronautical Sciences*, vol. 13, July, 1946, pp. 364–376.

12 "Design and Analog Computer Analysis of an Optimum Third-Order Nonlinear Servomechanism," by H. G. Doll and T. M. Stout, published in this issue, pp., 513–525.

13 Discussion, K. V. Diprose, "Automatic and Manual Control," Academic Press, New York, N. Y., 1952, p. 304.

14 "A Method for Solving Third and Higher-Order Nonlinear Differential Equations," by Y. H. Ku, *Journal of The Franklin Institute*, vol. 256, September, 1953, pp. 229–243.

15 "The Effects of Backlash and Speed Dependent Friction on the Stability of Closed-Cycle Control Systems," by A. Tustin, *Journal of The Institution of Electrical Engineers*, vol. 94, part IIA, May, 1947, pp. 143–151.

16 "On Some Nonlinear Phenomena in Regulatory Systems," by L. C. Goldfarb, *Avtomatika i Telemekhanika*, vol. 8, 1947, pp. 349–383, Translation in National Bureau of Standards Report 1691, May 29, 1952.

17 "Locus Methods for Regulators with Friction," by W. Oppelt, *Zeitschrift VDI*, vol. 90, June, 1948, pp. 179–183. Translation in National Bureau of Standards Report 1691, May 29, 1952.

18 "Theory of Relay Servomechanisms," by J. R. Dutilh, *Onde Electrique*, vol. 30, October, 1950, pp. 438–445.

19 "A Frequency Response Method for Analyzing and Synthe-

sizing Contactor Servomechanisms," by R. J. Kochenburger, Trans. AIEE, vol. 69, part 1, 1950, pp. 270–284.

20 "A Graphical Analysis for Nonlinear Systems," by P. S. Hsia, Proceedings of The Institution of Electrical Engineers, vol. 99, part II, April, 1952, pp. 125–134.

21 "How to Analyse Control Systems Graphically," by H. M. Paynter, *Control Engineering*, vol. 2, February, 1955, pp. 30–35, March, pp. 72–78.

22 "A Method of Analysing the Effect of Certain Kinds of Nonlinearity in Closed-Cycle Control Systems," by A. Tustin, *Journal of The Institution of Electrical Engineers*, vol. 94, part IIA, May, 1947, pp. 152–160.

23 "Number Series Method of Solving Linear and Nonlinear Differential Equations," by A. Madwed, Massachusetts Institute of Technology, Cambridge, Mass., Instrumentation Laboratory Report 6445-T-26, April, 1950.

24 "A Step-by-Step Method for Transient Analysis of Feedback Systems With One Nonlinear Element," by T. M. Stout, AIEE Technical Paper No. 56—779.

25 "An Analysis of Relay Servomechanisms," by D. A. Kahn, Trans. AIEE, vol. 68, part 2, Applications and Industry, 1949, pp. 1079–1088.

26 "Theory of Servomechanisms," by H. L. Hazen, *Journal of The Franklin Institute*, vol. 218, September, 1934, pp. 279–330.

27 "Backlash in a Velocity Lag Servomechanism," by N. B. Nichols, Trans. AIEE, vol. 73, part 2, Applications and Industry, 1954, pp. 462–467.

28 "Analysis of Nonlinear Servos by Phase-Plane-Delta Method," by R. N. Buland, *Journal of The Franklin Institute*, vol. 257, January, 1954, pp. 37–48.

29 "Oscillation of a Third-Order Nonlinear Autonomous System," by L. L. Rauch, "Contributions to the Theory of Nonlinear Oscillations," Princeton University Press, Princeton, N. J., 1950, pp. 39–88.

30 "Analysis and Design Principles of Second- and Higher-Order Saturating Servomechanisms," by R. E. Kalman, AIEE Paper 55–551.

31 "Sinusoidal Analysis of Feedback Control Systems Containing Nonlinear Elements," by E. C. Johnson, Trans. AIEE, vol. 71, part 2, Applications and Industry, 1952, pp. 169–181.

32 "Dissymmetrical Servomechanisms," by J. Loeb and J. D. Lebel, *Annales des Télécommunications*, vol. 9, October, 1954, pp. 282–286.

33 "The Mechanism of Subharmonic Generation in a Feedback System," by J. C. West and J. L. Douce, The Institution of Electrical Engineers Paper 1693, 1954.

34 "Nonlinear Control Systems with Random Inputs," by R. C. Booton, Transactions of the Institute of Radio Engineers—Professional Group on Circuit Theory, vol. CT-1, March, 1954, pp. 9–18.

35 "Operating Modes of a Servomechanism with Nonlinear Friction," by H. Lauer, *Journal of The Franklin Institute*, vol. 255, June, 1953, pp. 497–511.

36 "Phase-Plane Analysis of Automatic Control Systems With Nonlinear Gain Elements," by R. E. Kalman, Trans. AIEE, vol. 73, part 2, Applications and Industry, 1954, pp. 383–390.

Appendix 1

Phase-Plane Method

Construction of Trajectories

Direct Methods. Phase-plane trajectories for a second-order system can be obtained in a variety of ways. An x-y plotting table or oscilloscope might be connected to the actual system or an analog computer to record trajectories directly.

If the system is linear in some part of the plane, ordinary analytical methods can be used to solve the appropriate differential equations. Points representing values of error and error rate at a particular time may then be plotted and joined by a continuous curve. For example, the saturated region in the examples is governed by the equation

$$J\ddot{e} + f\dot{e} \pm T_m = 0 \qquad [15]$$

which has the solutions

$$e = e_0 + \frac{T_m}{f}t - \frac{T_m J}{f^2}\left(1 - e^{-\frac{f}{J}t}\right) \qquad [16]$$

$$\dot{e} = \frac{T_m}{f}\left(1 - e^{-\frac{f}{J}t}\right) \qquad [17]$$

when the initial error rate is zero and the negative sign is used in Equation [15]. In this case, it is comparatively simple to eliminate the time variable t analytically; the result is

$$e - e_0 = -\frac{T_m J}{f^2}\ln\left(1 - \frac{f}{T_m}\dot{e}\right) - \frac{J}{f}\dot{e} \qquad [18]$$

The trajectory obtained from Equation [18] by taking $e_0 = 0$, and a similar trajectory obtained by using the positive sign in Equation [15], can be shifted horizontally to pass through any point in the phase plane.

Indirect Analytical Methods. If these were the only methods for obtaining trajectories, the phase-plane would not be very valuable. Fortunately, however, trajectories can be constructed when the original differential equation cannot be solved directly for the variables as functions of time. As a simple illustration of an analytic method for doing this consider the differential equation applicable to the dead zone in Example 5

$$J\ddot{e} + f\dot{e} = 0 \qquad [19]$$

For convenience, we introduce the new variables

$$x = e \qquad [20]$$

$$y = \frac{dx}{dt} = \frac{de}{dt} = \dot{e} \qquad [21]$$

Since the error acceleration is $\ddot{e} = dy/dt$, Equation [19] may be written

$$\frac{dy}{dt} = -\frac{f}{J}y \qquad [22]$$

Dividing Equation [22] by Equation [21], we obtain

$$\frac{dy/dt}{dx/dt} = \frac{dy}{dx} = -\frac{f}{J} \qquad [23]$$

a first-order differential equation for the trajectories. Integration of Equation [23] gives

$$y = -\frac{f}{J}x + K_1 \qquad [24]$$

where K_1 is a constant of integration to be evaluated from initial conditions. With K_1 evaluated and the original notation restored, Equation [24] becomes

$$\dot{e} = -\frac{f}{J}(e - e_0) \qquad [25]$$

an algebraic expression for the trajectories in the dead zone.

In general, the original differential equation is likely to be more complicated than Equation [19] and may be written

$$\ddot{e} + \phi(e, \dot{e})\dot{e} + \psi(e, \dot{e})e = 0 \qquad [26]$$

The same process that produced Equation [23] now gives

$$\frac{dy}{dx} = \frac{-\phi(x, y)y - \psi(x, y)x}{y} \qquad [27]$$

Although Equation [27] is a first-order differential equation in x and y, it may not be solvable by analytic methods and graphical methods may therefore become necessary.

Isocline Method. One graphical method for solving Equation [27] involves the use of "isoclines" which are the loci of points of

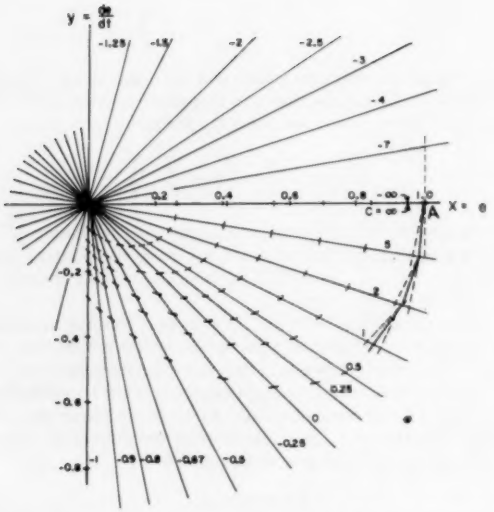


FIG. 14 TRAJECTORY CONSTRUCTION FOR LINEAR SYSTEM BY ISOCLINE METHOD

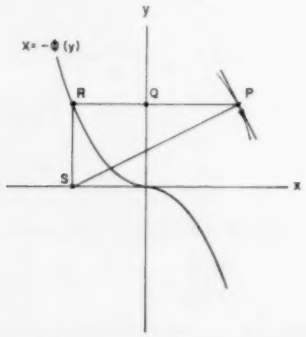


FIG. 15 BASIC PRINCIPLE OF LIENARD CONSTRUCTION

constant slope $dy/dx = C$. In the linear region, the isoclines are given by

$$\frac{-fy - Kx}{Jy} = C \qquad [28]$$

or, if $J = f = K = 1$, by

$$y = \frac{-1}{1 + C}x \qquad [29]$$

Isoclines plotted from Equation [29] are shown in Fig. 14. To construct a trajectory starting at some point A (say, $x = 1, y = 0$):

1 Draw a line through A with the indicated slope C (at this point $C = \infty$, so the line is vertical).

2 Draw a line through A with the slope associated with the next isocline (in this case, $C = 5$).

3 Draw a line bisecting the angle between the two lines drawn through A.

The bisector thus drawn is assumed to approximate the trajectory between the two isoclines. The intersection of this bisector with the second isocline gives a new initial point at which the process can be repeated. A few examples of the process are shown in Fig. 14.

The accuracy of this construction is increased by using a large number of segments to approximate a trajectory. If only moderate accuracy is required it is possible to superimpose short line segments of the proper slope on each isocline (as shown in the fourth quadrant of Fig. 14) and to sketch the trajectories freehand.

Lienard's Construction. Another graphical method, known as Lienard's construction (3), is available for differential equations which can be put into the form

$$\frac{dy}{dx} = \frac{-\phi(y) - x}{y} \qquad [30]$$

Before any trajectories are constructed, the curve $x = -\phi(y)$ is plotted in the phase plane, as shown in Fig. 15. To construct a trajectory starting at some point P:

1 Draw a line through P parallel to the x-axis, intersecting the curve at point R.

2 Draw a line from R to S perpendicular to the x-axis.

3 Draw the line PS.

4 Draw a line through P perpendicular to PS.

The line drawn through P in step 4 is tangent to the trajectory, since its slope is

$$dy/dx = -PR/RS \text{ and } PR = PQ + QR = x + \phi(y)$$
$$RS = y$$

In practice, steps 3 and 4 can be simplified by using a compass to construct an arc through P with its center at S; a short segment of this arc may be a good approximation of the trajectory if both have the same curvature. Applications of this method can be found in a paper by Lauer (35).

Time From Phase-Plane Plots

Although arrows on the trajectories show the direction of increasing time, the time associated with any particular point on the trajectory is not immediately available. Additional work is necessary to find, for example, the response time for a step input or the period of a sustained oscillation from trajectories constructed by any of the graphical methods. Two basic procedures are available for this purpose.

From the basic definition

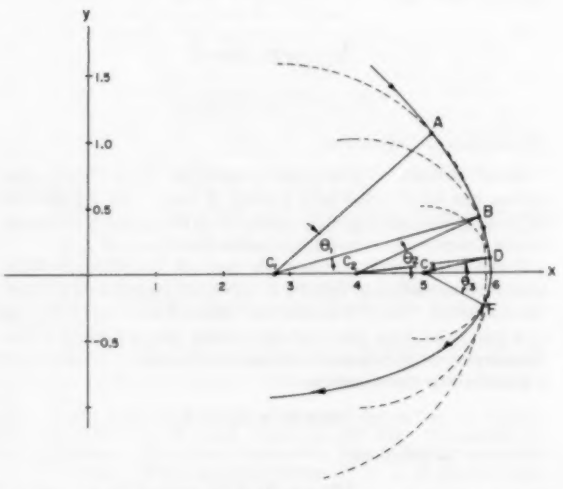


FIG. 16 APPROXIMATION OF TRAJECTORIES BY CIRCULAR ARCS FOR DETERMINING TIME INTERVALS

$$y = \frac{de}{dt} = \frac{dx}{dt} \quad [31]$$

it follows that

$$dt = \frac{dx}{y} \quad [32]$$

The time required to go from point A to point B on a trajectory is therefore

$$T_{AB} = \int_{x_B}^{x_A} \frac{dx}{y} \cong \sum_{A}^{B} \frac{\Delta x}{y_{avg}} \quad [33]$$

where y_{avg} is the average velocity for the distance Δx. This sort of calculation is somewhat inconvenient, especially if y goes through zero between A and B. Equation [33] may be useful in a qualitative way, however, for showing which of two trajectories spends the least time between A and B.

A more direct procedure, usable even when y goes to zero on the segment of trajectory under consideration, has been suggested by Diprose (13). His procedure is based on approximation of sections of the trajectory by arcs of circles centered on the x-axis and could be applied as shown in Fig. 16. The center C_1 of the first circle is located by trial and error so that arc AB closely approximates segment AB of the trajectory. The time interval between A and B is then given by

$$T_{AB} = \theta_1 \tau \quad [34]$$

where τ is the ratio of the scale factors on the x and y-axes, defined as

$$\tau = \frac{\text{number of } x \text{ units per division}}{\text{number of } y \text{ units per division}} \quad [35]$$

In the case shown, $\theta_1 = 26$ deg $= 0.454$ radians, and $\tau = 1.0/0.5 = 2$, so that

$$T_{AB} = (0.454)(2) = 0.908 \text{ sec}$$

In similar fashion, segments BD and DF are approximated by arcs of circles having their centers at C_2 and C_3, angles θ_2 and θ_3 are measured, and the corresponding times calculated.

Where detailed knowledge of the trajectories is not needed and a qualitative idea of system behavior is sufficient, the labor required for construction of trajectories and calculation of time intervals can be avoided by using the notion of singular points.

Singular Points

As may easily be imagined, difficulty in constructing system trajectories is experienced if dy/dx takes the form $0/0$, since the slope of the trajectory is then indeterminate. Such points are called "singular points" and are treated at length in texts on nonlinear mechanics (1 to 3).

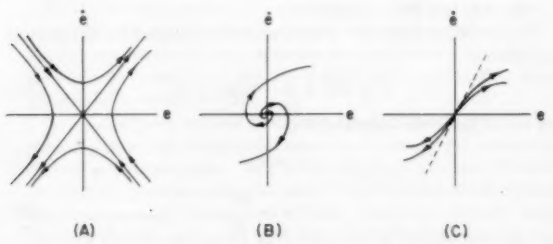


FIG. 17 UNSTABLE SINGULAR POINTS: (A) SADDLE POINT; (B) SPIRAL OR FOCAL POINT; (C) NODAL POINT

For our purposes, we will assume that Equation [27] has been put into the form

$$\frac{dy}{dx} = \frac{ax + by + F(x, y)}{y} \quad [36]$$

$$\cong \frac{ax + by}{y} \quad [37]$$

by expanding the numerator in a power series and then dropping the higher powers of x and y represented by $F(x, y)$. The singular points of this equation all occur at the origin. The nature of the singular points and the nearby trajectories depends on the coefficients a and b:

(1) $b^2 + 4a < 0$. The singular point is a *center* or *vortex* point if $b = 0$ and is surrounded by elliptical trajectories, none of which ever reaches the origin, as shown in Fig. 1(C). If $b \neq 0$, the singular point is a *spiral* or *focal point*, surrounded by spiral trajectories which all reach the origin, as shown in Fig. 1(A).

(2) $b^2 + 4a = 0$. The singular point is a *nodal point* or *node*. All trajectories reach the origin tangent to a single straight line, as shown in Fig. 1(B).

(3) $b^2 + 4a > 0$. The singular point is a *nodal point* or *node* if $a < 0$. If $a > 0$, the singular point is a *saddle point*, shown in Fig. 17(A). Two trajectories pass through a saddle point; the remainder approach the saddle point and turn away.

Spiral and nodal points are stable (trajectories approach the singular point) if $b < 0$ and unstable (trajectories leave the singular point) if $b > 0$. Unstable spiral and nodal points are shown in Figs. 17(B) and 17(C).

The saddle point is a point of unstable equilibrium and is encountered in control systems in which the torque-error relation is

$$T = T_m \sin e \quad [38]$$

as might occur with certain types of error-sensing devices. In this case, the singular points do not occur only at the origin. A phase-plane plot of typical behavior for such a system is given in Fig. 18.

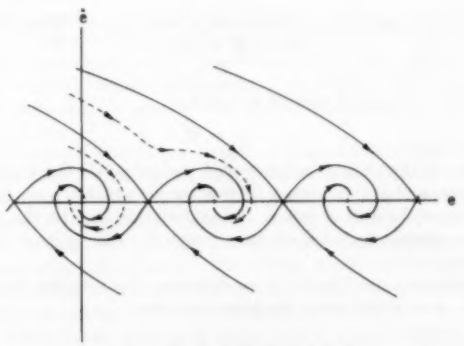


FIG. 18 PHASE-PLANE REPRESENTATION OF SYSTEM WITH SINUSOIDAL TORQUE-ERROR RELATION, SHOWING FOCAL POINTS AND SADDLE POINTS

Systems With Time-Varying Inputs

Ramp Inputs. In the examples considered so far, the input has been a step function. With the system at rest, the initial conditions are $e(0) = e_0$ and $\dot{e}(0) = 0$, a point on the e or x-axis of the phase plane. The graphical constructions which have been described are applicable, with appropriate modifications, to some systems subjected to ramp inputs.

As an example, we will consider a system described by the equation

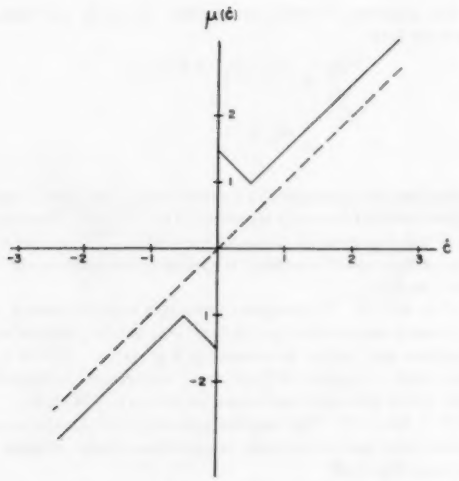


FIG. 19 FRICTION TORQUE VERSUS OUTPUT VELOCITY

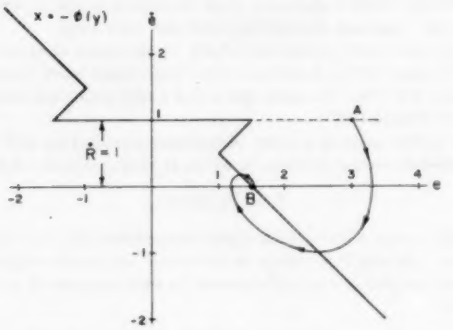


FIG. 20 PHASE-PLANE CONSTRUCTION FOR SYSTEM WITH RAMP INPUT ($\dot{R} = 1.0$)

$$\ddot{c} + \mu(\dot{c}) + c = r \quad [39]$$

$$= R + \dot{R}t \quad [40]$$

where c is the output or controlled variable, r is the input or reference, and $\mu(\dot{c})$ denotes a friction torque which is a nonlinear function of the output rate $\dot{c}$. As indicated in Fig. 19, the total friction torque includes viscous, coulomb, and friction components.

In order to use Lienard's construction, the equation for the system is rewritten using the basic definitions

$$e = r - c = R + \dot{R}t - c \quad [41]$$

$$\dot{e} = \dot{r} - \dot{c} = \dot{R} - \dot{c} \quad [42]$$

$$\ddot{e} = \ddot{r} - \ddot{c} = -\ddot{c} \quad [43]$$

giving

$$\ddot{e} + \mu(\dot{R} - \dot{e}) + e = 0 \quad [44]$$

Substituting $x = e$ and $y = \dot{e}$, Equation [44] can be written

$$\frac{dy}{dx} = \frac{-\mu(\dot{R} - y) - x}{y} \quad [45]$$

$$= \frac{-\phi(y) - x}{y} \quad [46]$$

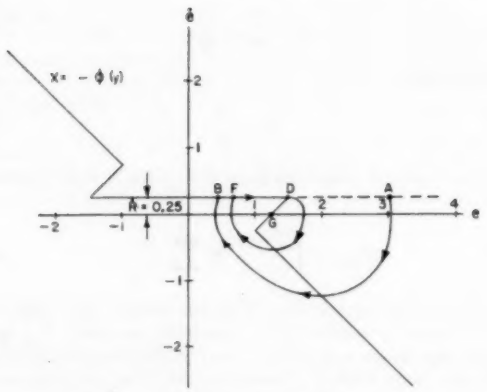


FIG. 21 PHASE-PLANE CONSTRUCTION FOR SYSTEM WITH RAMP INPUT ($\dot{R} = 0.25$)

where

$$\phi(y) = \mu(\dot{R} - y) \quad [47]$$

Equation [47] shows that the curve $x = -\phi(y)$ used in the Lienard construction is simply the friction curve $\mu(\dot{c})$ rotated 90 deg and then translated vertically along the y-axis to account for $\dot{R}$.

For $\dot{R} = 1$, the curve is translated upward one unit, as shown in Fig. 20. With $R = 3$ the trajectory starts (at A) with $c(0) = 3$, $\dot{c}(0) = 1$, and ends (at B) where $e_{ss} = 1.5$ and $\dot{e} = 0$. Point B may be called a stable focal point.

For $0 < \dot{R} < 0.5$, say $\dot{R} = 0.25$, the steady-state error in this system is not constant but periodic, as shown in Fig. 21. With $R = 3$, the trajectory starts at A with $e(0) = 3$, $\dot{e}(0) = 0.25$, and goes to B. According to Equation [46], points for which $x = -\phi(y)$ and $y \neq 0$ are points of zero slope, so the trajectory follows the curve from B to D. The trajectory leaves the curve at D, and returns to the curve at F where a periodic motion begins. The singular point at G is an unstable focal point.

General Inputs. If this system were subjected to a general time-varying input, say $r = r(t)$, Equation [39] would become

$$\ddot{c} + \mu(\dot{c}) + c = r(t) \quad [48]$$

Letting $x = c$ and $y = \dot{c}$, we would obtain the equation

$$\frac{dy}{dx} = \frac{r(t) - \mu(y) - x}{y} \quad [49]$$

which is similar to Equation [27] except for the numerator term $r(t)$. The presence of this term destroys the unique dependence of dy/dx on the phase-plane co-ordinates x and y. Curves can still be drawn showing the history of the system, using extensions of the methods already described, but the significance of the curves is no longer the same. Procedures for cases of this type are given by Ku (14) and Buland (28).

Higher-Order Systems. Faced with the third-order differential equation

$$\dddot{e} + a_2\ddot{e} + a_1\dot{e} + a_0 e = 0 \quad [50]$$

we could make the substitutions

$$x = e \quad [51]$$

$$y = \dot{e} = \frac{dx}{dt} \quad [52]$$

$$z = \ddot{e} = \frac{dy}{dt} \quad [53]$$

and examine system behavior in a three-dimensional phase space having co-ordinates x, y, and z. Because of difficulties in plotting curves in three dimensions, we would probably prefer to project the system trajectories into two planes, where the differential equations would become

$$\frac{dz}{dy} = \frac{-a_2 z - a_1 y - a_0 x}{z} \quad [54]$$

$$\frac{dy}{dx} = \frac{z}{y} \quad [55]$$

or

$$\frac{dx}{dy} = \frac{y}{z} \quad [55a]$$

Methods for constructing trajectories for third and higher-order systems have been presented by Ku (14), and some of the topological aspects of third-order systems (singular points and periodic solutions) are considered by Rauch (29). Phase-space considerations also have been used in the design of optimum relay servomechanisms (12).

Some higher-order systems are adequately approximated by second-order differential equations; implications of this fact for phase-plane analysis are examined in some detail by Kalman (36).

Appendix 2

Describing-Function Method

In the describing-function method as outlined in the body of this paper, it is assumed that:

1 The input to the nonlinear element is sinusoidal.
2 Only one nonlinear element occurs in the system.
3 The characteristics of the nonlinear element are independent of frequency.

Inability to satisfy these assumptions in a particular problem introduces difficulties which we propose to discuss briefly.

With a sinusoidal input, the output of the nonlinear element will necessarily contain harmonic components as well as the fundamental component used to define the describing function. Ordinarily, these harmonic components are smaller than the fundamental component and are, in addition, reduced in magnitude by the filtering action of dynamic elements in the system. Where the harmonic components are not negligible at the input to the nonlinear element, a more complete analysis along lines suggested by Johnson is required (31).

If the nonlinear element does not have a symmetrical input-output characteristic, the output will also contain a constant or d-c component. This component is not filtered in the usual control system and may, in fact, tend to be greatly amplified by integrators in the system. The magnitude of the d-c component at the output of a nonlinear element depends on the magnitude of both the d-c and sinusoidal components of the input, as does the magnitude of the fundamental component of the output. Equilibrium conditions in such systems are determined by consideration of two interrelated criteria similar to Equation [8]; details are available elsewhere (32).

Some nonlinear control systems may exhibit a peculiar type of response in which a sinusoidal input at frequency $n\omega$ produces an output at frequency ω. In a study of this phenomena, West and Douce (33) have introduced a describing function which expresses the fundamental component of the nonlinear element output (having the frequency ω) as a function of two sinusoidal input amplitudes, one with frequency ω and another with frequency $n\omega$. Only one equilibrium condition, similar to Equation [8], was important in their study, that relating to components at frequency ω.

The describing-function method is generally applied to systems having only one nonlinear element. If several nonlinearities occur simultaneously, it may be possible to represent them by a single describing function, as in the case of dead zone and saturation shown in Fig. 12. If two nonlinearities are separated by energy-storage elements (or if one nonlinearity is inextricably mixed with energy-storage elements), the combination must be represented by a describing function which is both amplitude- and frequency-dependent. The single curve $-1/G_D$ in the polar plot must be replaced by a family of curves for a number of frequencies, and the condition for a sustained oscillation becomes

$$G(\omega_{ss}) = \frac{-1}{G_D(X_i, \omega_{ss})} \quad [56]$$

where ω_{ss} denotes the frequency of oscillation.

A quasi-linearization method, similar in concept to the describing-function method, has been proposed by Booton for the study of systems subjected to random inputs (34). In his method, the nonlinear element is characterized by its response to a random input, and the result of the analysis is an estimate of the system r-m-s error.

Discussion

R. W. Bass.[4] This paper gives a clear summary of the methods now employed in nonlinear control-system analysis and synthesis.

In the writer's opinion, the describing-function method and (to a lesser extent) the phase-plane method are destined to be superseded by techniques better adapted to the nature of the problem.

The phase-plane, used in any of the forms mentioned by the author (Appendix 1), is a very powerful tool. An even more rapid method for obtaining a picture of the transient and steady-state response to steps and ramps has recently been introduced by R. E. Kalman.[5,6]

For systems of order $n > 2$, Kalman's techniques can be applied in the corresponding n-dimensional phase space. Such investigations[7] undoubtedly will continue to prove fruitful. However, they seem better adapted to analyses of a specific system (or class of systems) than to provide new synthesis insights (as they did for $n = 2, 3$).

The describing-function method is highly overrated as a synthesis tool. Although it can be rendered mathematically legitimate,[8] it provides at most a negative criterion (even when it works). It does often indicate correctly whether or not the system is (statically) stable, but of course it has nothing to offer concerning optimization. Worse still, it can lead to the appearance of auxiliary nonlinear phenomena which are undesirable but which the method does not reveal.[8]

For example, some relay servos cannot be stabilized by a compensating network unless so much error-rate feedback is employed that the servo "chatters" in response to a step input.

[4] Department of Mathematics, Princeton University, Princeton, N. J.

[5] "Physical and Mathematical Mechanisms of Instability in Nonlinear Automatic Control Systems," by R. E. Kalman, published in this issue, pp. 547–552.

[6] See Bibliography (36).

[7] "Analysis and Design Considerations of Second and Higher Order Saturating Servomechanisms," by R. E. Kalman, Trans. AIEE, vol. 74, part 2, Applications and Industry, 1955, no. 2, pp. 294–309.

[8] "Equivalent Linearization, Nonlinear Circuit Synthesis, and the Stabilization and Optimization of Control Systems," by R. W. Bass, to be published in Proceedings of the Symposium on Nonlinear Circuit Analysis, MRI Symposia Series, vol. 6, October, 1956.

This not only may degrade the response time, but (as the author has informed the writer) will cause the relay to wear out more quickly.

Proofs of these statements, and a discussion of some new alternative methods, are to be found elsewhere.[8]

HERBERT SAUNDERS.[9] The author has presented a most interesting and timely paper on this subject. As stated by the author, there exist numerous methods of solving nonlinear differential equations but none is completely satisfactory. For engineering purposes one needs some sort of solution even if it is only approximate. On many occasions a reduction or a great simplification will reduce the complexity of the differential equations employed in describing the problem so that the analysis may be simple as well as economical in respect to final results. Since most of our prevalent systems are of second order, many solutions are based on the linear second-order differential equation in which we may consider the higher-order equations as insignificant. If a significant nonlinearity which definitely cannot be ignored exists, the analytic solution becomes extremely difficult and arduous. The phase-plane method becomes very important when the equations cannot be integrated in closed form. Other writers[10–13] similarly have gone into great detail on the advantages and disadvantages of the phase plane; their work is not repeated here. The phase-plane delta method has been touched on slightly, yet is one of the most powerful phase-plane methods available. Actually, the Lienard method as discussed in the Appendix of the paper is but a special case of this method. Consider a general differential equation

$$J\ddot{e} + H(e, \dot{e}) = 0 \quad \ldots\ldots\ldots\ldots\ldots\ldots [57]$$

where $H(e, \dot{e})$ is a nonlinear function including time as well as describing any imposed external disturbance to which the system's response is required. Following Jacobsen,[11] the system can be put in a simpler form. Let

$$H(e, \dot{e}) = h(e, \dot{e}, t) + fe$$

and if

$$\omega^2 = f/J$$

then

$$\ddot{e} + \omega^2(e + \delta) = 0 \quad \ldots\ldots\ldots\ldots\ldots\ldots [58]$$

where

$$\delta = \frac{1}{f}\, g(e, \dot{e}, t)$$

If $\dot{e}/\omega = y$ then

$$\ddot{e} = \omega^2 y \frac{dy}{de}$$

Then Equation [58] simplifies to the following

$$y \frac{dy}{de} + e + \delta = 0 \quad \ldots\ldots\ldots\ldots\ldots\ldots [59]$$

or

$$\frac{dy}{de} = -\frac{e + \delta}{y}$$

Both Jacobsen[11] and Bishop[14] demonstrate the actual geometrical construction of this method. For further details the foregoing papers should be consulted. Considering that the equation

$$\ddot{e} + g(\dot{e}) + \omega^2 e = 0 \quad \ldots\ldots\ldots\ldots\ldots\ldots [60]$$

is but a special case of the phase-plane delta method, the graphical procedure is faster and more accurate than the Lienard method. This method can be applied to further problems concerned with systems subjected to transient loadings[15] and multidegree-of-freedom systems.[16]

AUTHOR'S CLOSURE

As workers in the field are well aware, there is an extensive literature on the subjects of nonlinear servomechanisms and mechanics. In writing this introductory paper, no attempt was made to survey the entire literature. The references provided by the discussers are a welcome addition to the paper. By tracking down the references listed in the references, and so on *ad infinitum*, the reader can find still additional information and can easily compile an extensive bibliography.

The author shares the feeling that existing methods leave something to be desired. The present methods are probably adequate for second-order systems whose nonlinearities can be concentrated in a single box and which are subjected to specialized inputs such as steps, ramps, or sinusoids. The more complicated problems presented by higher-order systems, containing several isolated nonlinearities and subjected to realistic inputs, are beginning to receive the attention they deserve. Research workers in the field can profitably focus their attention on these problems.

It may be well to add that analog or digital computers can, in principle, be used to study these complex systems. Computers can be regarded either as system models or mathematical machines, and their speed and flexibility permit systematic examination of a large number of alternative designs. Paper-and-pencil methods are, nevertheless, important. They provide insights into system behavior which cannot be obtained by looking at a pile of computer records. Experience with paper-and-pencil methods helps the designer to interpret computer results, diagnose troubles when they occur, and decide on a rational system change for the next test.

The goal, methods of analysis which answer all of the designer's questions with no effort on his part, is unobtainable. Any steps toward the goal will be a useful contribution to the field.

---

[9] Missiles and Ordnance Systems Department, General Electric Company, Philadelphia, Pa. Assoc. Mem. ASME.

[10] "Phase Plane Analysis of Nonlinear Control Systems," by I. R. Dalton, Research Report No. 3, University of Toronto, November, 1954.

[11] "On a General Method of Solving Second-Order Ordinary Differential Equations by Phase Plane Displacements," by L. S. Jacobsen, *Journal of Applied Mechanics*, Trans. ASME, vol. 74, 1952, pp. 543–553.

[12] "Geometrical Methods in the Analysis of Ordinary Differential Equations," by J. Kestin and S. K. Zaremba, *Applied Scientific Research*, section B, vol. 3, 1953, pp. 149–189.

[13] "On Motions of an Oscillatory System Under the Influence of Flip-Flop Controls," by I. Flügge-Lotz and K. Klotter, NACA TM 1237, November, 1949.

[14] "On the Graphical Solution of Transient Vibration Problems," by R. E. D. Bishop, Proceedings of The Institution of Mechanical Engineers, vol. 168, 1954, pp. 299–322.

[15] "Response of an Elastically Non-Linear System to Transient Disturbances," by R. L. Evaldson, R. S. Ayre, and L. S. Jacobsen, *Journal of The Franklin Institute*, vol. 248, December, 1949, pp. 473–494.

[16] "Transient Vibrations of Linear Multidegree-of-Freedom Systems by Phase-Plane Method," by R. S. Ayre, *Journal of The Franklin Institute*, vol. 253, no. 2, 1952, p. 153.

How to Obtain Describing Functions for Nonlinear Feedback Systems

By KARL KLOTTER,[1] STANFORD, CALIF.

A method is presented by which Describing Functions for nonlinear elements, whose behavior is defined by nonlinear differential equations (as contrasted with nonlinear relationships between the input and output variables themselves), can be obtained immediately (without first solving the differential equations). The method is applied to feedback systems containing one linear and one nonlinear element each described by a second-order differential equation.

Stating the Problem

In linear-system analysis the concept of frequency response, in the form of (direct or inverse) transfer functions, is well known; and it has been found to be an extremely useful tool. As a matter of fact, it is of such usefulness that the temptation to extend this basically linear concept into the domain of nonlinear systems has proved irresistible. If we discount some suggestions in earlier sources (1),[2] R. Kochenburger (2, 3) is to be credited with having developed a nonlinear counterpart to the transfer function that became known as the "Describing Function."

To emphasize the analogy to the respective linear concept, in this paper we frequently will replace the term Describing Function by "equivalent transfer function" with qualifying adjectives, such as "direct," "inverse," "complex," "real," and so on, as the case may be. [Perhaps a more appropriate (although still longer) term would be "equivalent frequency-response function" with the pertinent qualifying adjectives.]

Whatever name may be used for that function, the concept embodies the following idea:

In a nonlinear element a sinusoidal input will produce a nonsinusoidal, although (for steady state) periodic, output. In order to perpetuate the use of a (complex) amplitude ratio one may work with the first harmonic of the periodic output.

This suggestion sounds reasonable enough and good arguments can be offered to support it (4). Such arguments are essentially based on the filtering effect which takes place in the linear part of a transmitting system, in particular of a closed-loop system. On the other hand, because the higher harmonics, which are definitely present in the output, are disregarded, the results of using the describing function have to be watched attentively, even with some suspicion, and they always stand in need of careful checking and circumspect interpretation. In fact, the whole procedure warrants some more basic justification or at least some exploring of its inherent limitations. In this paper, however, we will not concern ourselves with those questions of justification or limitation. We will accept, more or less naively, the suggestion as proposed and the concept as customarily used and applied.

In looking over the applications of the Describing-Function concept which have been made to date, one cannot fail to note that all cases which have been treated, as dead band (7, 8), saturation (7), limiting (3), linkages (5), contactors (2), coulomb friction (6), are such cases where the relationship between input and output can be given by a nonlinear expression for the variables themselves as contrasted with a differential equation. The treatment of the Describing Function in (9) is further testimony to that fact.

Where, very occasionally, an element has been considered whose relationship between input and output is expressed by a differential equation, the only suggestion offered has been to solve the differential equation by some computing technique (numerical or instrumental) and then to determine the first Fourier component of that computed output. Such a procedure is extremely tedious and clumsy, however.

There was, up to date, no practical way known for obtaining the Describing Function for elements possessing differential relationships and hence for feedback systems containing such elements. It is the purpose of this paper to present a convenient method for producing the desired information, the first harmonic of the output function of an element whose behavior is described by a nonlinear differential equation.

In earlier papers (10, 11) the author has explained the fundamentals of such a method. Therefore here we will dispense with explaining the philosophy underlying it; we will rather focus our efforts on showing how the method can be used for producing the Describing Function of a nonlinear element without actually solving its differential equation. It will become clear, the author hopes, that the method is well suited for the purpose at hand; in fact, it seems to be custom-made for the present needs. And, certainly, its potentialities have not yet been exhausted.

A Method for Producing the Describing Function (Equivalent Transfer Function) of a Nonlinear Element

Let

$$E(z) \equiv M(z) - \kappa^2 y(t) = 0 \quad \ldots\ldots\ldots\ldots [1]$$

denote the nonlinear differential equation of the element under consideration with $y(t) = Y \cos \omega t$ standing for the input and z for the unknown output. By way of illustration we will treat the differential equation

$$E \equiv \ddot{z} + 2D\kappa g(\dot{z}) + \kappa^2 f(z) - \kappa^2 Y \cos \omega t = 0 \ldots\ldots [2]$$

which happens to be of second order. It contains two nonlinear terms $g(\dot{z})$ and $f(z)$. For simplicity of presentation we will assume here that these two functions $g(\dot{z})$ and $f(z)$ are odd functions of their respective arguments. [In paper (11) the treatment is outlined also for the case of nonodd functions.] And as an example of Equation [2] we are going to treat the system described by

$$g(\dot{z}) = \dot{z} \qquad f(z) = z(1 + \mu^2 z^2)$$

hence

$$E \equiv \ddot{z} + 2D\kappa\dot{z} + \kappa^2 z(1 + \mu^2 z^2) - \kappa^2 Y \cos \omega t = 0 \ldots [3]$$

Equation [3] is known as "Duffing's differential equation."

[1] Professor of Engineering Mechanics, Stanford University. Mem. ASME.

[2] Numbers in parentheses refer to the Bibliography at the end of the paper.

Contributed by the Instruments and Regulators Division of The American Society of Mechanical Engineers and presented at the ASME-AIEE Conference on Nonlinear Control Systems, Princeton, N. J., March 26–28, 1956.

Note: Statements and opinions advanced in papers are to be understood as individual expressions of their authors and not those of the Society. Manuscript received at ASME Headquarters, January 17, 1956. Paper No. 56—IRD-5.

According to what was said earlier we wish to produce an approximation

$$\tilde{z} = Z \cos(\omega t - \epsilon) \quad [4]$$

to the output function $z(t)$, with the parameters Z and ϵ determined in such a way that they represent "best" values in some sense or other.

Because the function $\tilde{z}(t)$ Equation [4] cannot satisfy the differential Equation [2] at every instant, we are going to satisfy it in some "weighted average." Appropriate weight functions can be shown (10, 11) to be $\cos \omega t$ and $\sin \omega t$. The method (known as the Ritz averaging method or Ritz-Galerkin method) then asks for

$$\left.\begin{aligned} \int_0^{2\pi} E[\tilde{z}(\sigma)] \cos \sigma d\sigma = 0 \\ \text{and} \\ \int_0^{2\pi} E[\tilde{z}(\sigma)] \sin \sigma d\sigma = 0 \end{aligned}\right\} \quad [5]$$

These are two equations for determining the two parameters Z and ϵ in the Approximating Function [4].

Without giving here any justification of the procedure we just list the steps which are to be taken:

Step 1. From the two functions $f(z)$ and $g(\dot{z})$, which appear in the differential Equation [2] of the element under consideration we derive two new functions $F(Z)$ and $G(Z, \omega)$ by performing the following integrations

$$\left.\begin{aligned} F(Z) = \frac{4}{\pi} \frac{1}{Z} \int_0^{\pi/2} f(Z \cos \sigma) \cos \sigma d\sigma \\ \text{or equivalently} \\ = \frac{4}{\pi} \frac{1}{Z} \int_0^{\pi/2} f(Z \sin \sigma) \sin \sigma d\sigma \end{aligned}\right\} \quad [6a]$$

and

$$\left.\begin{aligned} G(Z, \omega) = \frac{4}{\pi} \frac{1}{\kappa} \frac{1}{Z} \int_0^{\pi/2} g(Z\omega \sin \sigma) \sin \sigma d\sigma \\ \text{or equivalently} \\ = \frac{4}{\pi} \frac{1}{\kappa} \frac{1}{Z} \int_0^{\pi/2} g(Z\omega \cos \sigma) \cos \sigma d\sigma \end{aligned}\right\} \quad [6b]$$

By way of explanation we may add that Equations [6] amount to the following procedure: Introduce $Z \cos \omega t$ and $Z\omega \sin \omega t$ for z and $\dot{z}$, respectively, replace the functions f and g of these arguments by functions of multiple arguments ($2\omega t$, $3\omega t$, etc.), and retain the fundamental terms only.

Step 2. Using those functions F and G and abbreviating ω/κ by η we obtain the amplitude Z and the phase shift ϵ from the two equations (which follow from Equations [5])

$$[F - \eta^2]^2 + 4D^2G^2 = \left(\frac{Y}{Z}\right)^2 \quad [7a]$$

and

$$\tan \epsilon = \frac{2DG}{F - \eta^2} \quad [7b]$$

Passing on, now, to the special differential Equation [3] we find F and G to be

$$F = 1 + \frac{3}{4} Z^2 \quad \text{and} \quad G = \eta \quad [8]$$

where use is made of the abbreviation $Z = \mu Z$. Hence the Equations [7] read

$$\left[1 + \frac{3}{4} Z^2 - \eta^2\right]^2 + 4D^2\eta^2 = \left(\frac{Y}{Z}\right)^2 \quad [9a]$$

and

$$\tan \epsilon = \frac{2D\eta}{1 + \frac{3}{4} Z^2 - \eta^2} \quad [9b]$$

Equations [9] give both the modulus $|N| = Y/Z$ and the argument ϵ of the "inverse equivalent transfer function"

$$N = \left(\frac{Y}{Z}\right) e^{i\epsilon} \quad [10]$$

This (complex) transfer function can be plotted as a family of curves, either with Z as a parameter of the family and η varying along the individual curves, or with η as a parameter of the family and Z varying along the individual curves. In the next section, Fig. 2, we will see the plot for N used for dealing with feedback systems containing the element which is under consideration here. The solid curves of one set (parabolas) in Fig. 2 represent the loci of N for values of $Z = 0$; 1; 2; 3, respectively; the other set (straight lines) represents the loci of N for the parameter values $\eta = 0$; 1; 2; 3, respectively.

One realizes readily that Equations [7] for the general case of differential Equation [2] or Equations [9] for the special case of differential Equation [3] are equivalent to

$$|N| \cos \epsilon = F - \eta^2$$

$$|N| \sin \epsilon = 2DG$$

with

$$\frac{Y}{Z} = |N| = \sqrt{(F - \eta^2)^2 + 4D^2G^2}$$

or

$$|N| \cos \epsilon = 1 + \frac{3}{4} Z^2 - \eta^2$$

$$|N| \sin \epsilon = 2D\eta$$

with

$$\frac{Y}{Z} = |N| = \sqrt{\left(1 + \frac{3}{4} Z^2 - \eta^2\right)^2 + 4D^2\eta^2}$$

respectively.

Application of the Equivalent Transfer Function to Feedback Systems

In parts (a) through (d) of Fig. 1 four cases of feedback systems are listed, each containing two elements. The first element I, is supposed to be linear in all cases, the second one II, either linear or nonlinear. The (complex) inverse transfer function of the linear element I is denoted by L_I; the (complex) inverse transfer function of element II, if linear, by L_{II}, if nonlinear by N_{II} (then to be called "equivalent" transfer function).

First, consider all elements to be linear. Then the over-all (complex) direct transfer functions are (with asterisks denoting complex amplitudes):

In cases (*a*) and (*b*)

$$\frac{Z^*}{U^*} = \frac{1}{1 + L_I L_{II}} \quad \ldots\ldots [11a, b]$$

In case (*c*)

$$\frac{Y^*}{U^*} = \frac{L_{II}}{1 + L_I L_{II}} \quad \ldots\ldots [11c]$$

In case (*d*)

$$\frac{Y^*}{U^*} = \frac{L_I}{1 + L_I L_{II}} \quad \ldots\ldots [11d]$$

Self-sustained oscillations in all of these systems will occur when the denominator vanishes

$$1 + L_I L_{II} = 0 \quad \ldots\ldots [12]$$

Equation [12] is the basis for the use of the Nyquist criterion.

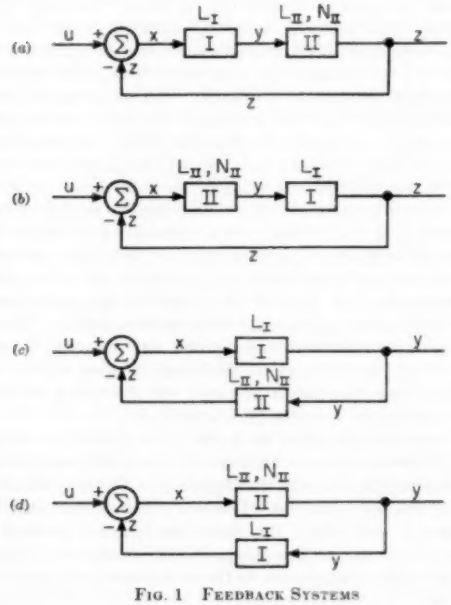
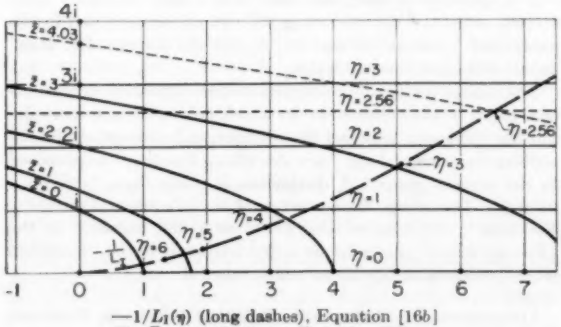


FIG. 1 FEEDBACK SYSTEMS



$-1/L_I(\eta)$ (long dashes), Equation [16*b*]
$N_{II}(Z, \eta)$ (solid lines), Equation [19]

FIG. 2 PLOT OF FUNCTIONS

Next, consider element II to be nonlinear in all cases. The only change we have to make is to replace the inverse transfer function L_{II} by its nonlinear counterpart, the inverse equivalent transfer function N_{II}, as developed in the preceding section. Hence self-sustained oscillations may be expected if there are roots to the equation

$$1 + L_I N_{II} = 0 \quad \ldots\ldots [13a]$$

or

$$N_{II} = -1/L_I \quad \ldots\ldots [13b]$$

By way of example we consider such loops where the behavior of both elements is described by a second-order differential equation. The differential equation for the linear element I will be specified as

$$\ddot{y} + \delta_1 \dot{y} = \kappa_1^2 x \quad \text{or} \quad \ddot{z} + \delta_1 \dot{z} = \kappa_1^2 y \quad \ldots\ldots [14]$$

the differential equation for element II, if nonlinear, will be assumed identical with Equation [3] (Duffing's equation), specifically

$$\left.\begin{aligned} &\ddot{z} + \delta_2 \dot{z} + \kappa_2^2 z(1 + \mu^2 z^2) = \kappa_2^2 y \\ \text{or} \\ &\ddot{y} + \delta_2 \dot{y} + \kappa_2^2 y(1 + \mu^2 y^2) = \kappa_2^2 x \end{aligned}\right\} \quad \ldots\ldots [15]$$

If both elements were linear ($\mu = 0$), their (complex) inverse transfer functions would be

$$L_I = [-\eta^2 + 2D_1 \eta i] \frac{\kappa_2^2}{\kappa_1^2} \quad \ldots\ldots [16a]$$

hence

$$-1/L_I = \frac{\kappa_1^2}{\kappa_2^2} \frac{\eta^2 + 2D_1 \eta i}{\eta^4 + 4D_1^2 \eta^2} \quad \ldots\ldots [16b]$$

and

$$L_{II} = (1 - \eta^2) + 2D_2 \eta i \quad \ldots\ldots [17]$$

where

$$\eta = \omega/\kappa_2, \qquad 2D_1 = \delta_1/\kappa_2, \qquad 2D_2 = \delta_2/\kappa_2 \quad \ldots\ldots [18]$$

In the nonlinear case L_{II} is to be replaced by N_{II}

$$N_{II} = \left(1 + \frac{3}{4} Z^2 - \eta^2\right) + 2D_2 \eta i \quad \ldots\ldots [19]$$

In Fig. 2 plots are shown of N_{II} (solid lines), Equation [19], and $-1/L_I$ (long dashes), Equation [16*b*], with the following numerical values used

$$\left.\begin{aligned} \kappa_1^2 &= 5000/\text{sec}^2 & \kappa_2^2 &= 100/\text{sec}^2 \\ \delta_1 &= 10/\text{sec} & \delta_2 &= 10/\text{sec} \end{aligned}\right\} \quad \ldots\ldots [20]$$

As is indicated on the plot, there does exist an intersection of the curve $-1/L_I(\eta)$ with the family of curves $N_{II}(Z, \eta)$ for a common parameter value η. For the example shown this value happens to be $\eta = 2.56$. Hence (stable) self-sustained oscillations will occur. They have an (angular) frequency $\omega = 25.6$/sec and an amplitude Z determined by $\mu Z = 4.03$ (μ has not been specified here).

Just in passing it should be mentioned that if element II is linear ($\mu = 0$) the closed-loop systems may or may not be stable, depending on the values of the parameters. For the numerical values specified, the linear systems are unstable. This fact can be

verified easily either from the Nyquist criterion or from the Routh-Hurwitz criterion.

The differential equation of the closed-loop linear systems reads

$$z^{IV} + (\delta_1 + \delta_2)\ddot{z} + (\delta_1\delta_2 + \kappa_2^2)\dot{z} + \delta_1\kappa_2^2\dot{z} + \kappa_1^2\kappa_2^2 z = \kappa_1^2\kappa_2^2 u \quad . . [21]$$

Therefore, the Hurwitz determinants become in turn

$$\left.\begin{aligned} H_1 &= \delta_1 + \delta_2 \\ H_2 &= \delta_1^2\delta_2 + \delta_1\delta_2^2 + \delta_2\kappa_2^2 \\ H_3 &= \kappa_2^2[\delta_1\delta_2(\delta_1^2 + \delta_1\delta_2 + \kappa_2^2 - 2\kappa_2^2) - \kappa_1^2(\delta_1^2 + \delta_2^2)] \end{aligned}\right\} . . [22]$$

Whereas H_1 and H_2 are always positive (for positive damping) H_3 may or may not be positive. For the numerical values specified, H_3 is negative; hence the linear systems are unstable.

We have, therefore, before us systems which, if linear, prove to be unstable (showing indefinitely increasing amplitudes) but which can be restrained to a limit cycle by the addition of a nonlinear term, $\mu^2 z^3$ in our case.

Details of the various dependencies (of amplitudes and phase angles on frequency ω, on damping coefficients δ_1 and δ_2, on nonlinearity μ^2, and so on) can be discussed; the results may be presented in a separate paper. Replacing the graphical procedure used here by an analytical one will help to improve the accuracy of the resulting numerical values.

Acknowledgments

This paper presents some partial results obtained in the course of studies aided by the National Science Foundation. The author wishes to acknowledge the assistance of Mr. Hugh L. Smith, graduate student in Engineering Mechanics at Stanford University.

BIBLIOGRAPHY

1 "Dynamik selbsttätiger Regelungen," by R. C. Oldenbourg and H. Sartorius, München, Germany, 1944, translated by H. I. Mason, "Dynamics of Automatic Control," New York, N. Y., 1948.

For further references to earlier papers see the discussion remarks by G. A. Philbrick to paper (2), and the remarks in (9), p. 556, concerning papers by L. C. Goldfarb, W. Oppelt, and A. Tustin.

2 "A Frequency Response Method for Analyzing and Synthesizing Contactor Servomechanisms," by R. J. Kochenburger, AIEE Trans., vol. 69, part 1, 1950, pp. 270–284.

3 "Limiting in Feedback Control Systems," by R. J. Kochenburger, AIEE Trans., vol. 72, part 2, July, 1953, pp. 180–194.

4 "Recent Advances in Nonlinear Servo Theory," by J. M. Loeb, Trans. ASME, vol. 76, 1954, pp. 1281–1290 (more literature is listed there).

5 "Sinusoidal Analysis of Feedback-Control Systems Containing Nonlinear Elements," by E. C. Johnson, AIEE Trans., vol. 71, part 2, July, 1952, pp. 169–181.

6 "Coulomb Friction in Feedback Control Systems," by V. B. Haas, Jr., AIEE Trans., vol. 72, part 2, May, 1953, pp. 119–126.

7 "Approximate Frequency-Response Methods for Representing Saturation and Dead Band," by H. Chestnut, Trans. ASME, vol. 76, 1954, pp. 1345–1364.

8 "Stability Characteristics of Closed-Loop Systems With Dead Band," by C. H. Thomas, Trans. ASME, vol. 76, 1954, pp. 1365–1382.

9 "Automatic Feedback Control System Synthesis," by J. C. Truxal, McGraw-Hill Book Company, Inc., New York, N. Y., 1955, chapter 10.

10 "Nonlinear Vibration Problems Treated by the Averaging Method of W. Ritz," by K. Klotter, Proceedings of the First National Congress of Applied Mechanics 1951, ASME, New York, N. Y., 1952, pp. 125–131.

11 "Steady State Vibrations in Systems Having Arbitrary Restoring and Arbitrary Damping Forces," by K. Klotter, Proceedings, Symposium Nonlinear Circuit Analysis, Polytechnic Institute of Brooklyn, Brooklyn, N. Y., 1953, pp. 234–257.

Discussion

R. W. Bass.[3] This paper greatly enlarges the class of feedback systems to which the describing-function method can be applied conveniently. The extension of the "equivalent transfer function" from elements described by nonlinear functions to elements described by nonlinear differential equations is both a realistic and a welcome advance in the art.

As the author says, "the whole procedure warrants (1) some more basic justification, or at least (2) some exploring of its inherent limitations."

The procedure is essentially a heuristic criterion for the existence of periodic solutions of nonlinear differential equations. A rigorous criterion as suggested in (1) would involve the solution of infinitely many implicit equations in infinitely many unknowns. For that there are two techniques available: (*a*) The Ritz-Galerkin variational methods, as developed; e.g., in Lichtenstein's book on nonlinear integral equations and in the recent work of Rothe; (*b*) the fixed-point techniques of Lefschetz and Leray-Schauder.

In a recent contribution to the Symposium on "Nonlinear Circuit Analysis,"[4] the writer investigated problems (1) and (2).

By using technique (*b*) the infinite system can be reduced to a single pair of simultaneous equations (which determine the frequency and the fundamental amplitude). These equations are a generalization of the classical bifurcation equations; if they can be solved there will be a self-sustained oscillation. In some cases they can be solved; this leads to an answer to question (1).

In general, however, it is too difficult to treat these equations directly. But on putting a certain term equal to zero, a pair of easily solved, approximate bifurcation equations is obtained. If one splits the author's Equations [12] or [13*b*] into real and imaginary parts, one finds just these equations. In other words, the describing-function method is actually a graphical technique for solving the approximate bifurcation equations. There are very general conditions under which the existence of this approximate solution necessitates the existence of an exact solution. When these conditions are met, the describing-function method is legitimate; question (1) is answered.

But it is often impractical to verify these conditions. What then of (2)? It can be shown that the higher the frequency found from Equations [12] and [13], the closer will be these equations to the true bifurcation equations. It seems plausible then that for high-frequency oscillations the describing-function method is more likely to be valid. For this and other reasons we suggest this as a tentative contribution to the author's second question.

Author's Closure

It is pleasing to learn that Dr. Bass is able to report some decisive steps toward answering the two questions posed in the paper and listed as (1) and (2) in the discussion. Dr. Bass' paper[4] will merit close attention.

The author perhaps is permitted to draw attention here to his own paper in the symposium[4] in which he discusses in more detail the difference between the conventional concept of the describing function and the "new describing function" as proposed in the present paper. A distinction is made there between a "Fourier Describing Function" and a "Hamilton Describing Function"; the former being based on a first harmonic in the "Fourier sense," the latter on a first harmonic in the "Hamilton sense" minimizing the Hamiltonian integral.

---

[3] Department of Mathematics, Princeton University, Princeton, N. J.

[4] Proceedings, Symposium on Nonlinear Circuit Analysis, Polytechnic Institute of Brooklyn, Brooklyn, N. Y., vol. 6, October, 1956 (to be published).

Design and Analog-Computer Analysis of an Optimum Third-Order Nonlinear Servomechanism

By H. G. DOLL[1] and T. M. STOUT[2]

Great interest has been shown recently in deliberately nonlinear control systems, including a class of programmed control systems. The design objective which distinguishes these systems is minimization of the response time for step inputs by proper automatic timing of relay operations. For a second-order relay servomechanism, the required switching relation is expressed by a curve in a two-dimensional phase plane which can be realized by a one-variable function generator in combination with linear elements. For a third-order system, the switching relation is expressed by a surface in a three-dimensional phase space and requires a two-variable function generator for its realization. The switching surface is computed in this paper for an idealized third-order positioning servomechanism having an output member characterized completely by its moment of inertia and a torque which varies linearly with time between two limits; small-signal nonlinearities such as backlash or relay threshold are neglected. An electro-optical two-variable function generator is employed in an analog-computer study of this system. For purposes of comparison, two alternative modes of control also are examined. The system using programmed control shows the expected superiority for step inputs, the advantage being greatest for small step magnitudes. For sinusoidal or random inputs, programmed control is superior when the input amplitude and/or frequency are low but exhibits some anomalous behavior for large amplitudes and/or frequencies, which result in inferior performance. Parameter tolerances for programmed control systems are somewhat less severe than anticipated.

Introduction

METHODS for the analysis and synthesis of linear feedback control systems have had an intensive development since 1940. At the same time, there has been a growing interest in the analysis and design of nonlinear systems, directed toward eliminating or minimizing harmful effects of nonlinearity and exploiting useful effects of nonlinearity. Work in the latter area has been concerned particularly with the design of systems subject to saturation, leading to the development of a class of deliberately saturated or relay control systems to be described. This development is motivated by the following line of reasoning.

In many control systems the correcting action is made proportional to a linear combination of the system error, error rate, integrated error, and possibly other signals. This proportionality cannot, however, exist under all conditions, since a maximum corrective action is always imposed by inherent limitations of the materials and energy sources employed. If these limits are to be reached only under extreme conditions, the system evidently must be designed to operate most of the time at less than maximum capacity. Accuracy and speed of response, it is felt, are therefore not as good as they might be.

In the limit, as the proportionality constant is increased, the saturating linear system becomes effectively a relay system, in which a manipulated variable (such as field voltage) is switched between two fixed values. Stable operation can be achieved if the switching operation depends on a sufficient number of variables.

If the switching operation is made a nonlinear function of the system variables, a relay system can be designed which returns to equilibrium from any combination of initial values of the variables in a minimum time. In particular, the system is optimum in the sense that its response time for a step input is a minimum; this is the usual basis of design of such systems. Relay systems designed to obtain this objective may be called "programmed control systems" since the controller automatically programs or times the switching operations.

Work on a programmed controller for a third-order system was done by H. G. Doll in 1942, resulting in a patent application (1)[3] in 1943; in this work the second-order system was considered a limiting special case of a third-order system. According to West (2), independent work on programmed control for a second-order positioning servomechanism is contained in an unpublished report written by F. C. Williams in 1942. Oldenburger states (13) that he worked out the design for second and third-order cases in 1944. Additional work on second-order systems has been reported by McDonald (3), Hopkin (4), Uttley and Hammond (5), Lathrop (6), West (7), Bushaw (8), Neiswander and MacNeal (9), and others. Research on third-order systems was recently described by Bogner (10), Silva (11), Preston (12), and Hopkin and Iwama (25).

The possibility of extending the programmed control idea to higher-order systems has been recognized but no experimental work in this direction has been reported. Practical difficulties arise from the fact that a programmed controller for an nth-order system requires at least an $(n - 1)$-variable-function generator. Lack of suitable two-variable function generators has hampered even laboratory investigation of third-order systems. In the most extensive experimental investigation of a third-order system yet published (12), Preston used two single-variable diode function generators and an interpolating servomechanism to approximate the required function of two variables.

After a brief review of the theory underlying a second-order

[1] Schlumberger Instrument Company, Ridgefield, Conn.

[2] Work done at Schlumberger Instrument Company, Ridgefield, Conn.; now with the Ramo-Wooldridge Corporation, Los Angeles, California.

Contributed by the Instruments and Regulators Division of The American Society of Mechanical Engineers and presented at the ASME-AIEE Conference on Nonlinear Control Systems, Princeton, N. J., March 26–28, 1956.

Note: Statements and opinions advanced in papers are to be understood as individual expressions of their authors and not those of the Society. Manuscript received at ASME Headquarters, January 6, 1956. Paper No. 56—IRD-10.

[3] Numbers in parentheses refer to the Bibliography at the end of the paper.

programmed control system, this paper describes the calculation of the switching relation for a particular third-order system and a laboratory realization, using an electro-optical two-variable function generator, for use in an analog-computer investigation. Results of computer tests with a variety of inputs are given for this system as well as competitive linear and relay modes of control.

Second-Order System

In the positioning servomechanism considered in past studies, it generally has been assumed that the output member was completely characterized by its moment of inertia J and that the torque supplied by the motor ($T = \pm T_m$) could be reversed instantaneously. Parasitic nonlinearities, such as backlash or relay hysteresis, have ordinarily been neglected. With these assumptions, the behavior of the system is described by the second-order differential equation

$$J\ddot{c} = T = \pm T_m \quad [1]$$

where c is the controlled variable or output, and dots denote differentiation with respect to time. For step inputs, Equation [1] becomes

$$J\ddot{\epsilon} = -T = \mp T_m \quad [2]$$

Optimum step-function response is obtained by using maximum accelerating torque until the error is reduced to half its original value and maximum decelerating torque until the error is reduced to zero, as shown in Fig. 1.

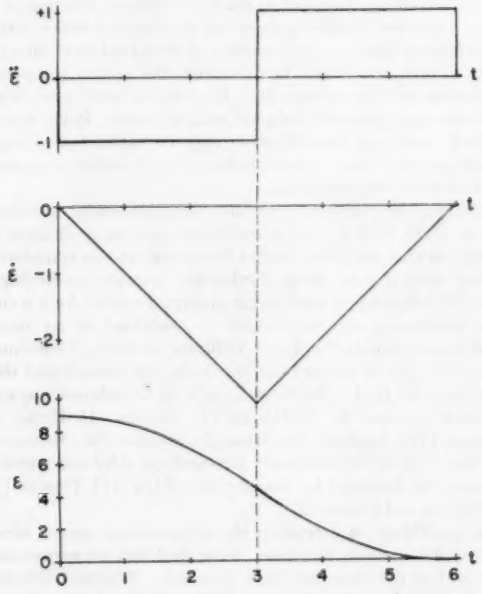


Fig. 1 Step Response of a Programmed Second-Order Servomechanism

To obtain optimum response for all step magnitudes, the torque is determined continuously from the relation

$$T = T_m \operatorname{sgn}\left[\frac{2T_m}{J} e + \dot{e}|\dot{e}|\right] \quad [3]$$

where sgn x denotes $x/|x|$.

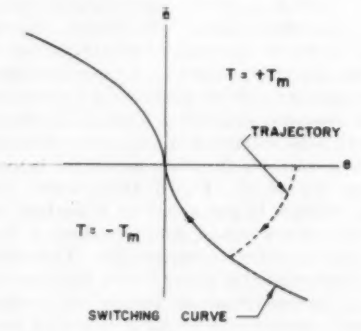
Torque reversal occurs on the switching curve defined by

$$\frac{2T_m}{J} e + \dot{e}|\dot{e}| = 0 \quad [4]$$

which divides the e-$\dot{e}$ phase plane into regions of positive and negative torque as shown in Fig. 2(A). With this switching relation, all system trajectories have the same general appearance and coincide with the switching curve approaching the origin. The response time for step inputs is

$$t_r = 2\sqrt{\frac{J}{T_m}}\, e_0 \quad [5]$$

where e_0 is the initial error or step magnitude (14).



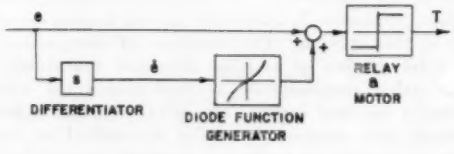


Fig. 2 Switching Curve for a Programmed Second-Order Servomechanism and a Possible Realization

A possible physical realization of the switching relation is shown schematically in Fig. 2(B). As indicated, the sense of the torque is a function of both error and error rate, but the physical realization requires only a single-variable function generator in combination with linear elements. Negative output rate ($-\dot{c}$) can be used instead of error rate without affecting the response for step inputs; this substitution does, however, result in a steady-state error for ramp inputs proportional to the square of the input rate.

The position and shape of the switching curve can be changed to obtain an optimum step response in the face of continuous coulomb or viscous friction torques, neither of which affect the order of the differential equation (14).

Third-Order System

A finite time ordinarily will be required for the torque to go from one extreme value to the other. The variation of torque with time may take several forms, all of which raise the order of the differential equation and require a more complicated controller.

Bogner (10) and Preston (12) consider systems in which the torque varies exponentially with time, described by the equations

$$J\ddot{c} = T \quad [6]$$

$$\tau_f \dot{T} + T = \pm T_m \quad [7]$$

which reduce to

$$\tau_f \dddot{c} + \ddot{c} = \pm \frac{T_m}{J} \quad [8]$$

The time constant τ_f might represent the field time constant of a d-c servomotor.

In the system considered here, the torque is assumed to vary linearly with time between its extreme values, so that system operation is described by the equations

$$J\ddot{c} = T \quad [6]$$

where

$$\dot{T} = \pm \frac{2T_m}{\tau} \qquad |T| < T_m$$

$$T = \pm T_m \qquad \text{otherwise} \quad [9]$$

and τ is the time required for the torque to go from one extreme value to the other. These equations, which may also be written

$$\dddot{c} = \pm \frac{2T_m}{\tau J} \qquad |T| < T_m \quad [10]$$

$$\ddot{c} = \pm \frac{T_m}{J} \qquad \text{otherwise} \quad [11]$$

can be regarded as a description of a highly idealized steering system in which a rudder is driven between limit stops by a constant-speed motor, producing a torque on the steered body proportional to the rudder angle (1),[4] (15). They also may be taken as a simplified description of a missile or aircraft roll-control system in which a rate-limited hydraulic servo is used to position the control surfaces (16–18).

Step-Function Response. Figs. 3 and 4 show two characteristic step-function responses of this third-order system. For small step inputs, such that the initial error e_0 satisfies the relation

$$e_0 \leq \frac{T_m}{2J} \tau^2 \quad [12]$$

the maximum error acceleration or torque is not required and the response takes the form shown in Fig. 3. Under these conditions the response time can be shown to be

$$t_r = \sqrt[3]{\frac{16\,J\tau}{T_m}}\, e_0 \quad [13]$$

when t_r is the time required for the error to become zero.

For larger step inputs, maximum error acceleration is reached and the response is as shown in Fig. 4. For very large step inputs, such that maximum torque is employed for almost the entire response, the response times are only slightly greater than the values given by Equation [5].

Figs. 1 and 4 are drawn for the same T_m, J, and response time, and show a smaller initial error for the third-order system; the explanation is, of course, that the acceleration in the third-order system is at most equal to that of the second-order system, and the error rate is smaller at all times. The third-order system therefore has a greater response time for the same initial error or step magnitude. Nevertheless, the responses shown in Fig. 3 and 4 are the best possible with the prescribed values of J, T_m, and dT/dt; there is no switching procedure which will return the system to equilibrium in a shorter time. The response time can be reduced only by decreasing J, increasing T_m or dT/dt, or by some other modification of the system (such as a brake added only during the decelerating interval).

[4] In reference (1), the switching procedure and mechanism are developed for a system whose equation of motion also includes a term corresponding to the restoring torque produced by aerodynamic forces acting on the steered body itself.

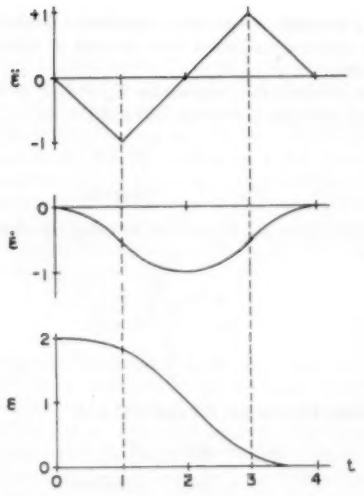


Fig. 3 Step Response of a Programmed Third-Order System

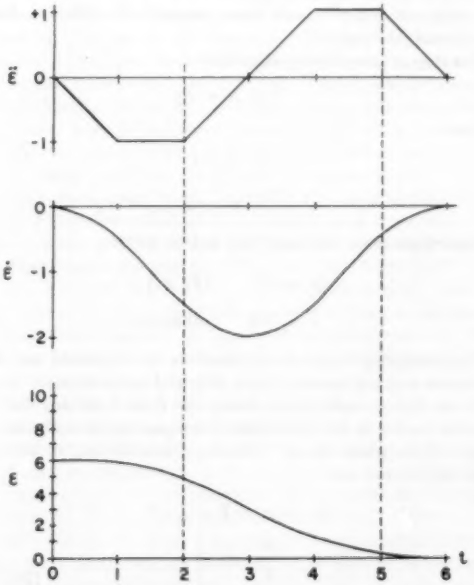


Fig. 4 Step Response of a Programmed Third-Order System

The torque rate dT/dt is reversed twice, as indicated by the dashed lines in Figs. 3 and 4. Since these switching operations occur at unique combinations of e, $\dot{e}$, and $\ddot{e}$, the switching relation for this third-order system is expressed by a "switching surface" in a three-dimensional phase space.[5]

Calculation of Switching Surface. The switching surface can be obtained by associating with each point in the e-$\dot{e}$ plane the value of error acceleration required if the system is executing an optimum response. This process is best carried out by con-

[5] Reference (16) describes means for proper timing of only the first reversal which result in near-optimum performance.

structing a number of optimum trajectories, starting from the origin and working backward from the end of the trajectory in negative time.

For ease of calculation, the values $T_m = 1$, $J = 1$, and $\tau = 2$ will be used, reducing Equations [10] and [11] to

$$\dddot{c} = \pm 1 \qquad |T| < 1 \ldots\ldots [14]$$

$$\ddot{c} = \pm 1 \qquad \text{otherwise} \ldots\ldots [15]$$

These substitutions are equivalent to using the nondimensional variables

$$\mathbf{c} = \frac{4J}{\tau^3 T_m} c \ldots\ldots [16]$$

$$\mathbf{t} = \frac{2}{\tau} t \ldots\ldots [17]$$

which convert Equations [10] and [11] into

$$\mathbf{c}''' = \pm 1 \qquad |T| < T_m \ldots\ldots [18]$$

$$\mathbf{c}'' = \pm 1 \qquad \text{otherwise} \ldots\ldots [19]$$

where primes denote differentiation with respect to $\mathbf{t}$. If Equations [14] and [15] are used, the discussion can be carried out in the original notation and later generalized with the help of Equations [16] and [17].

For step or ramp inputs, such that

$$r = R + \dot{R}t \ldots\ldots [20]$$

we have

$$\ddot{c} = -\ddot{e} \ldots\ldots [21]$$

$$\dddot{c} = -\dddot{e} \ldots\ldots [22]$$

so that Equations [14] and [15] can be written

$$\dddot{e} = \mp 1 \qquad |T| < 1 \ldots\ldots [23]$$

$$\ddot{e} = \mp 1 \qquad \text{otherwise} \ldots\ldots [24]$$

The switching relation can therefore be expressed in e-$\dot{e}$-$\ddot{e}$ coordinates and will serve for both step and ramp inputs.

As in the second-order system, the final switching operation occurs on a system trajectory passing through the origin of the phase space. The equations for one of two such trajectories are

$$\left.\begin{aligned} \ddot{e} &= -t = -T \\ \dot{e} &= -\frac{t^2}{2} \\ e &= -\frac{t^3}{6} \end{aligned}\right\} \ldots\ldots [25]$$

with t negative; $\ddot{e}$ and e are positive, T and $\dot{e}$ negative, and the projection of this trajectory appears in the fourth quadrant of the e-$\dot{e}$ plane. The other trajectory is symmetrical with respect to the origin.

At the final switching point $t = -t_s$ and

$$\left.\begin{aligned} \ddot{e} &= t_s \\ \dot{e} &= \frac{-t_s^2}{2} \qquad (t_s > 0) \\ e &= \frac{t_s^3}{6} \end{aligned}\right\} \ldots\ldots [26]$$

Half of the switching surface consists of trajectories terminating on those of Equation [25] at the co-ordinates given by Equation [26]; the other half is obtained by symmetry considerations. For $t \leq -t_s$, the equations are

$$\left.\begin{aligned} \ddot{e} &= t + 2t_s \\ \dot{e} &= \frac{t^2}{2} + 2t_s t + t_s^2 \\ &= \frac{1}{2}(t + 2t_s)^2 - t_s^2 \\ e &= \frac{t^3}{6} + t_s t^2 + t_s^2 t + \frac{t_s^3}{6} \\ &= \frac{1}{6}\{(t + 2t_s)[(t + 2t_s)^2 - 6t_s^2] + 6t_s^3\} \end{aligned}\right\} \ldots [27]$$

Letting $x = e$, $y = \dot{e}$, and $z = \ddot{e}$, eliminating t_s and rearranging in a form suitable for numerical computation, the equation of the switching surface is found to be

$$x = z\left(y - \frac{z^2}{3}\right) + \left(\frac{z^2}{2} - y\right)^{3/2} \ldots\ldots [28]$$

Although Equations [27] and [28] seemingly apply only when t_s is less than unity and therefore only when the step response is as shown in Fig. 3, they actually have a wider application. In the present system, the useful portion of the switching surface is bounded by the planes $\ddot{e} = z = \pm 1$, since $J = T_m = 1$ and the absolute value of $\ddot{e}$ cannot exceed unity. However, the shape of this portion of the surface is unaffected by the presence of a torque limit and would be unchanged if T_m were infinite. Hence quation [28] can be used to compute the entire switching surface. The planes $\ddot{e} = z = \pm 1$ are not to be considered part of the switching surface.

The torque limit must, of course, be considered in computing system trajectories; these trajectories are confined to the space between and including the bounding planes and may have, in some cases, segments in common with the intersection of the switching surface and the bounding planes. Equations [14] and

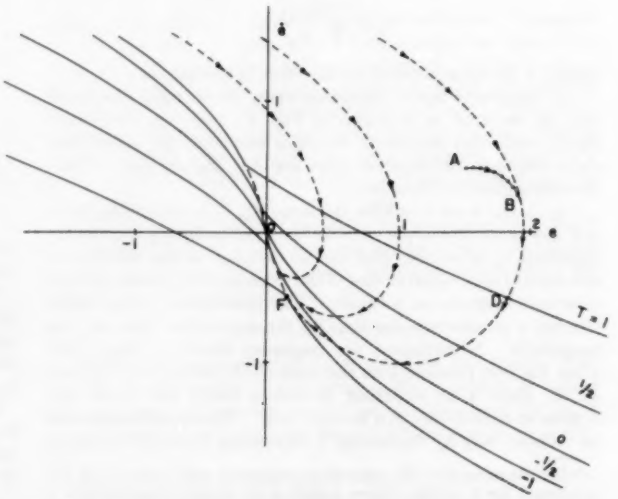


FIG. 5 SWITCHING SURFACE AND TRAJECTORIES FOR THE PROGRAMMED THIRD-ORDER SYSTEM

[15] are used alternately as the system goes from saturation in torque rate to saturation in torque magnitude, solutions being fitted together at the transition points. Typical trajectories computed in this manner are shown in Fig. 5, superimposed on a plot of the switching surface. Since $\ddot{e} = -T$ for the numerical values used here, the surface is plotted as lines of constant torque rather than error acceleration.

Controller Realization. Equations [6] and [21] state that

$$\ddot{c} = -\ddot{e} = \frac{T}{J} \quad [29]$$

for step and ramp inputs, indicating that $\ddot{e} = z$ in Equation [28] could be replaced by $-\ddot{e}$ (or $-T$, if $J = 1$) without detriment to the step or ramp response of the system. In the hypothetical steering or roll-control systems mentioned previously, the torque was assumed proportional to a control-surface deflection which could be measured by a position-sensing device (1). Alternatively $\ddot{c}$ might be measured directly; this method is advantageous if the system is subject to torque disturbances.

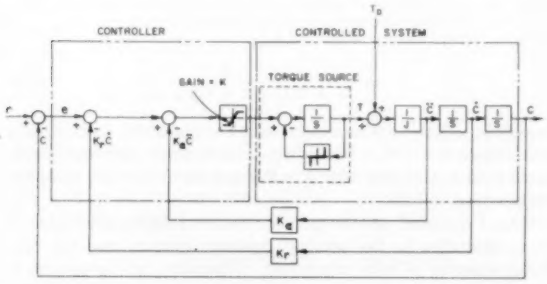


FIG. 6 BLOCK DIAGRAM OF THE PROGRAMMED THIRD-ORDER SYSTEM

If the torque is measured, the realization of the controller can take the form shown in Fig. 6. The output of the two-variable function generator is the torque computed from Equation [28], $T_c = -Jz = f(e, \dot{e})$, which must exist for each combination of e and $\dot{e}$ on an optimum trajectory. The relay input

$$\epsilon = T_c - T \quad [30]$$

the difference between the desired and actual torques, initiates action bringing the two torques into correspondence. Since the rate of change of torque called for on the optimum trajectories does not exceed the capabilities of the system, the controller can enforce continuous correspondence between the computed and actual torque after they have been brought into agreement.

Operation of the system under general conditions is described by

$$\dot{T} = \text{sgn}\,[f(e, \dot{e}) - T] \qquad |T| < 1 \quad [31]$$

$$T = \pm 1 \qquad \text{otherwise}$$

and can be understood with the help of Fig. 5. Assume that the system starts at point A with $T = -\ddot{e} = 0$. Since this point lies below the switching surface, the relay acts to make $dT/dt = +1$ and the system starts along the dotted path to point B. The initial slope of the trajectory is zero because, under all conditions

$$\frac{d\dot{e}}{de} = \frac{\ddot{e}}{\dot{e}} \quad [32]$$

and $\ddot{e} = 0$, $\dot{e} \neq 0$ at point A. At point B, the actual torque agrees with the torque used in computing the optimum trajectory ($T = 1$); the two trajectories are tangent at this point by virtue of Equation [32].

If the system is not subjected to further disturbances, the system follows the previously computed optimum trajectory from B to the origin. The first switching occurs at point D when the trajectory reaches the switching surface; here dT/dt becomes -1 and the torque starts from $+1$ toward -1. For this particular set of initial conditions, the second switching occurs (at point F) just as the torque reaches -1. For initial points farther from the origin, the trajectory follows the curve $T = -1$ before reaching F; for points closer to the origin, the second switching occurs on the trajectory joining F and O.

SYSTEMS STUDIED

An analog-computer investigation of this system, using a programmed controller as well as competitive linear and relay controllers, was conducted to examine the behavior for inputs which are not easily handled by analytical methods and to get an idea of the comparative performance of the various modes of control. Step-function response curves, obtained by analytical and graphical methods, were used to check the performance of the computer.

Programmed Controller. The electro-optical two-variable function generator used for programmed control of this system in these experiments employs a photoelectric tube to count pulses generated by the passage of lines on a rotating disk through a light beam. The light beam is positioned by a galvanometer excited by one input, while the duration of the counting interval is controlled by the other input. The number of lines counted per revolution is converted to an analog voltage which is a function of the two inputs. The nature of the function generator affects the analog-computer configuration in several ways.

The continuous variation of T over the switching surface is replaced by a stepwise variation using a number of steps which is limited by space and resolution considerations. In these tests, sixteen lines (corresponding to $T/T_m = +15/16, +13/16, +11/16, \ldots, +1/16, -1/16, \ldots, -15/16$) were used. Use of an even number of lines gives a region of zero torque on either side of the nominal zero-torque line, at the origin as well as in the rest of the plane. Some inaccuracy in the determination of switching points results from this approximation.

To make more efficient use of the available disk area and to avoid resolution problems in the neighborhood of the origin, it was considered advisable to express the switching surface in terms of new co-ordinates. A 45-deg rotation was arbitrarily adopted and $(e + \dot{e})$, $(e - \dot{e})$ were used as the inputs to the function generator. The range of values used was $-0.7 < (e + \dot{e}) < +0.7$ and $-1 < (e - \dot{e}) < +1$; operation was thereby limited to step inputs producing an initial error of approximately one unit.

The disk rotates at 1800 rpm in this function generator, so that samples of the output may be considered to be available every 1/30 sec. The theoretical upper frequency limit on the output of the function generator is therefore 15 cps, with a smaller limit (3 — 4 cps) imposed by practical filtering problems. To avoid possible difficulties introduced by filter lags, the analog computer was operated with a time unit $(\tau/2)$ of 10 sec.

A photograph of the disk and a diagram of the computer configuration, based directly on Fig. 6, are given in Figs. 7 and 8.

Linear Controller. A schematic diagram of a linear controller for this system is shown in Fig. 9. For the same numerical values as before and using acceleration feedback instead of torque feedback, operation of this system is governed by the equations

$$\dot{T} = K(e - K_r\dot{e} - K_a\ddot{c}) \quad [33]$$

FIG. 7 DISK USED IN FUNCTION GENERATOR

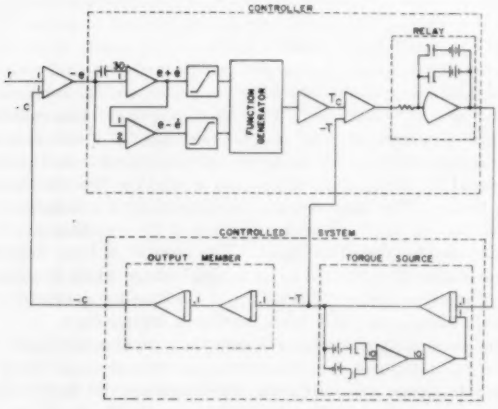


FIG. 8 COMPUTER CONFIGURATION FOR PROGRAMMED CONTROL SYSTEM

$$\dot{T} = \text{sgn}\,[e - K_r\dot{c} - K_a\ddot{c}] \quad \ldots\ldots [34]$$

$$T = \pm 1 \quad \ldots\ldots [35]$$

Equation [33] applies only if

$$|e - K_r\dot{c} - K_a\ddot{c}| < \frac{1}{K} \quad \ldots\ldots [36]$$

and

$$|T| < 1 \quad \ldots\ldots [37]$$

while Equation [34] applies if the inequality of Equation [36] is not satisfied, and Equation [35] applies if neither Equation [36] nor Equation [37] is satisfied.

In the region of linear operation described by Equation [33], the system transfer function is

$$\frac{C(s)}{R(s)} = \frac{K}{s^3 + K_aKs^2 + K_rK_s + K} \quad \ldots\ldots [38]$$

The gain K influences the speed of response or band width of the system and also fixes the size of the linear operating region. With $K = 1$, the value which was arbitrarily adopted, a unit step input with the system at rest will just produce saturation in $\ddot{c}$.

Values of the rate and acceleration constants K_r and K_a were selected, using data given by Graham and Lathrop (19), for linear behavior which was a satisfactory compromise between

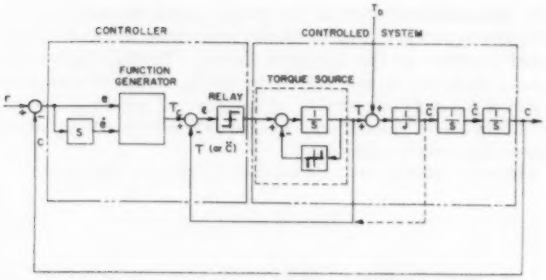


FIG. 9 BLOCK DIAGRAM OF THIRD-ORDER SYSTEM WITH LINEAR OR RELAY CONTROLLER

speed of response and degree of damping; the values used were $K_a = 1.9$ and $K_r = 2.28$. With these values, the transfer function becomes

$$\frac{C(s)}{R(s)} = \frac{1}{s^3 + 1.9s^2 + 2.28s + 1} = \frac{1}{(s + 0.7)(s^2 + 1.2\,s + 1.44)} \quad \ldots\ldots [39]$$

Transient solutions thus consist of an exponential term with a time constant $1/0.7 = 1.44$ units and damped sine and cosine terms with a damping ratio $\zeta = 0.5$ and an undamped resonant frequency $\omega_0 = 1.20$.

Relay Controller. If the gain K becomes infinite, the region of linear operation in the previous system vanishes and the controller becomes a relay controller. Operation of the system is now described by Equations [34] and [35]. For step inputs when $\dot{c} = -\dot{e}$, Equation [34] may be written

$$\left.\begin{aligned} T &= \text{sgn}\,[e + K_r\dot{e} - K_aT] \\ &= \text{sgn}\left[\frac{1}{K_a}e + \frac{K_r}{K_a}\dot{e} - T\right] \\ &= \text{sgn}\,[T_c - T] \end{aligned}\right\} \quad \ldots\ldots [40]$$

where

$$T_c = \frac{1}{K_a}e + \frac{K_r}{K_a}\dot{e} \quad \ldots\ldots [41]$$

Comparison of Equation [40] with Equation [31] indicates that the relay controller may be regarded not only as a linear controller with infinite gain but also as an approximation to the programmed controller, in which the curved surface is replaced by a plane. This observation suggests a basis for selection of the constants K_a and K_r, which should be given different values than were used in the linear controller.

From Equation [41], it is evident that the line $T_c = +1$ intersects the co-ordinate axes at

$$\left.\begin{aligned} E &= K_a \\ \dot{E} &= \frac{K_a}{K_r} \end{aligned}\right\} \quad \ldots\ldots [42]$$

Inspection of the switching surface in Fig. 5 suggests that suitable values would be found in the range

$$\left.\begin{aligned} 0.5 < E < 0.8 \\ 0.4 < \dot{E} < 0.5 \end{aligned}\right\} \quad \ldots\ldots [43]$$

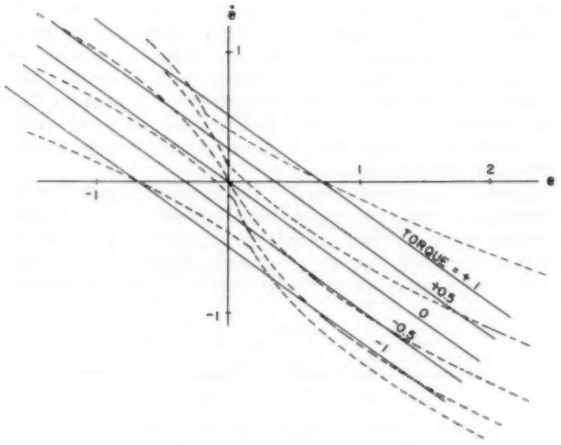


FIG. 10 APPROXIMATION OF SWITCHING SURFACE USED IN RELAY CONTROLLER FOR THIRD-ORDER SYSTEM

$$\left.\begin{aligned} 0.5 < K_a < 0.8 \\ 1 < K_r < 2 \end{aligned}\right\} \quad \ldots\ldots\ldots\ldots\ldots [44]$$

By experiment, the values $K_a = 0.7$ and $K_r = 1.4$ were chosen as a compromise between speed of response and degree of damping for both small and large step inputs. The resulting approximation of the switching surface is shown in Fig. 10.

RESULTS OF COMPUTER INVESTIGATION

A variety of inputs and disturbances was employed in the computer investigation. Results of these tests are summarized in the following sections.

Step Inputs. Responses of the three systems for step inputs, traced from the original records, are shown in Fig. 11. Response time, arbitrarily defined as the time required for the error to become and remain less than 0.025 unit, is plotted for the three systems in Fig. 12. For the smallest step magnitude employed, the response time for the programmed system is less than one half the response time for the relay system and about one fourth as great as for the linear system. For large step magnitudes, the differences are less pronounced.

Sinusoidal Inputs. Samples from more extensive records of tests using sinusoidal inputs are shown in Fig. 13; numerical values of the error magnitudes can be obtained from Fig. 14.

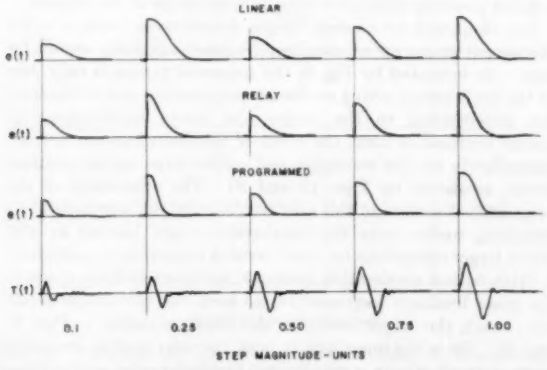


FIG. 11 STEP RESPONSE OF THIRD-ORDER SYSTEM WITH LINEAR, RELAY, AND PROGRAMMED CONTROL

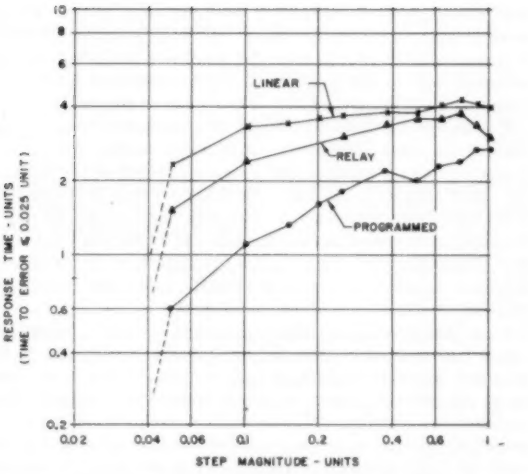


FIG. 12 RESPONSE TIME FOR STEP INPUTS WITH LINEAR, RELAY, AND PROGRAMMED CONTROL

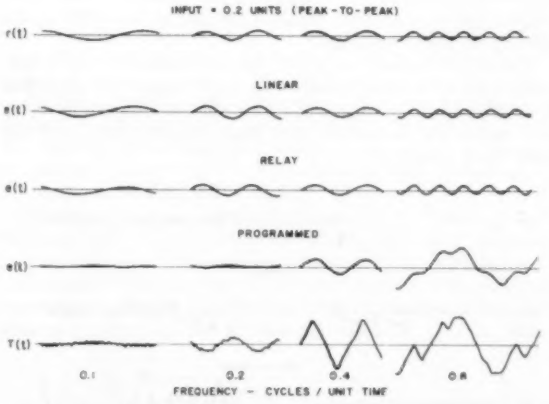


FIG. 13 RESPONSE OF THIRD-ORDER SYSTEM TO SINUSOIDAL INPUTS WITH LINEAR, RELAY, AND PROGRAMMED CONTROL

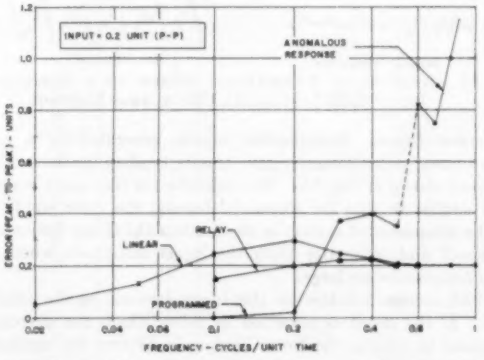


FIG. 14 ERROR MAGNITUDE FOR A SINUSOIDAL INPUT OF FIXED AMPLITUDE AND VARYING FREQUENCY

Figs. 13 and 14 indicate that the programmed system has a very small error at low frequencies and that its band width, defined arbitrarily as the frequency at which the error magni-

tude is one quarter or one half of the input magnitude, is several times greater than for the linear and relay systems. At high frequencies, however, the error for the linear and relay systems becomes equal to the input but the error magnitude for the programmed system becomes four or five times as great as the input.

The anomalous behavior of the programmed system at high input frequencies consists of fairly large torque and output motions having a fundamental frequency which is one third of the input frequency or less. As other records (not included here) show, this frequency is relatively constant as the input frequency is increased beyond the point at which this behavior is first observed. The error, being the difference of input and output, contains components at both the input and output frequencies.

Since the response of the programmed system to sinusoidal inputs had not been calculated analytically in advance of the computer experiments, this anomalous behavior was not anticipated. A limited number of checks indicated that the system equations are satisfied with reasonable accuracy in this mode of operation, so that no gross errors in the computer simulation could be held responsible. Subharmonic behavior in simpler nonlinear systems is discussed extensively in the literature on nonlinear mechanics [see, for example, Minorsky (20) or the more recent text by Stoker (21)] and isolated occurrences in feedback-control systems have been reported (22, 23), so that these results are not considered surprising.

The input frequency at which the anomalous behavior is first observed is an inverse function of the input magnitude, increasing from about 0.5 cycles/unit time for the case shown to about 0.75 cycles/unit time if the input magnitude is reduced to 0.025 units (peak-to-peak).

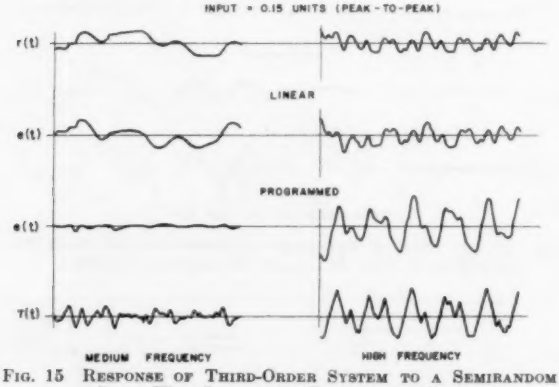


FIG. 15 RESPONSE OF THIRD-ORDER SYSTEM TO A SEMIRANDOM INPUT WITH LINEAR AND PROGRAMMED CONTROL

Random Inputs. Semirandom inputs, generated by a cam-driven linear potentiometer, gave results typified by the sections of record shown in Fig. 15. The behavior for this input is somewhat similar to that for sinusoidal inputs; the error amplitude for the programmed system is small when the input frequencies are small and is greater than the input amplitude when the input frequencies are large.

Which system is better for this input depends on the point of view. If the input is regarded as noise which the system is supposed to ignore, the linear system is better; the similarity of the input and error records indicates that the output is essentially stationary. If the input is regarded as a signal which the system should follow, the programmed system does a better job when the input frequencies are low and a poorer job otherwise.

A few tests with superimposed ramp and random inputs, or sinusoids of different frequencies and amplitudes, showed that the output of the programmed system can be predicted in a qualitative way by a superposition of the responses to the separate input components. This behavior might be explained by the fact that a feedback system is more nearly linear than the elements in the forward part of the loop (24). It suggests that the programmed system might give very inferior performance if subjected to a slowly varying input with superimposed high-frequency noise components.

Impulse Torque Disturbances. The effect of impulse torque disturbances which impart an initial velocity to the output member, correspondingly roughly to the effect of gusts on an airplane or missile, were also examined. Sections of the computer records are shown in Fig. 16, and data from the original

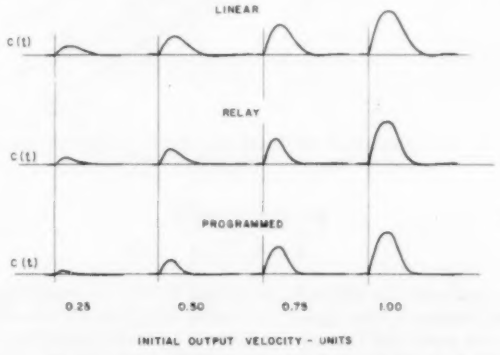


FIG. 16 RESPONSE OF THIRD-ORDER SYSTEM TO IMPULSE TORQUE DISTURBANCE WITH LINEAR, RELAY, AND PROGRAMMED CONTROL

records are summarized in Figs. 17 and 18. The programmed system has a three or four-to-one superiority in response time and maximum output magnitude for the smallest disturbance used and a smaller advantage for large disturbances.

For the programmed system, these records also indicate the error which would result from application of a ramp input. This is not true for the other systems which used output rate rather than error rate in the controller.

Sinusoidal Torque Disturbances. For step inputs, it is immaterial in principle whether the error acceleration is obtained by differentiating the error or by measuring the output acceleration or torque. In practice, however, there is a considerable difference between the three possibilities for other inputs. Double differentiation of the error signal with attendant noise problems (13) is avoided by use of output acceleration or torque; this also reduces problems associated with differentiation of step inputs.

For sinusoidal or random torque disturbances (with $r = 0$), the output and error acceleration are interchangeable except for sign. As indicated by Fig. 6, the measured torque is only part of the total torque acting on the output member and is therefore not proportional to the output (or error) acceleration. If torque feedback is used, the effect of the disturbance is not felt immediately by the controller and rather large output motions result, as shown by Figs. 19 and 20. The superiority of the programmed system in this case results from the steepness of its switching surface near the phase-space origin, leading to relatively larger corrections for small error or error-rate magnitudes.

With output acceleration feedback, as shown in Figs. 6 and 9, the inner feedback loop operates to keep the net torque small. As a result, the output motion is also small, as shown by Figs. 21 and 22. Since the inner loop in both the relay and programmed systems is effectively a rate-limited first-order relay servomechanism, it can be expected to operate almost perfectly for sinus-

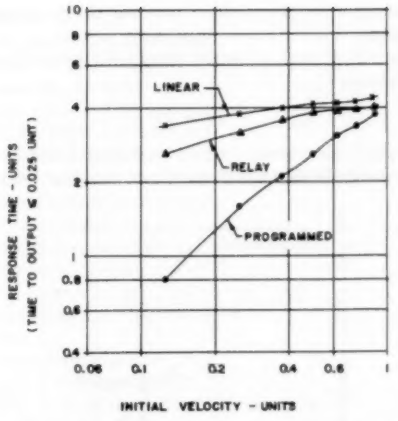


FIG. 17 RESPONSE TIME FOR IMPULSE TORQUE DISTURBANCES

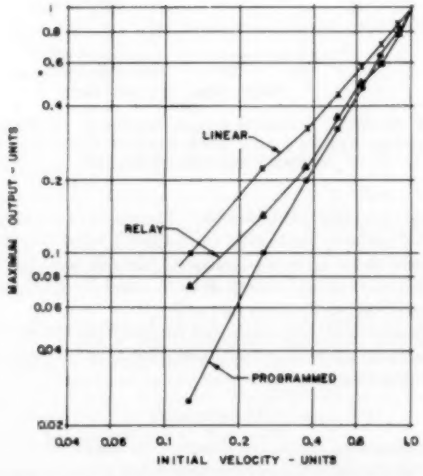


FIG. 18 MAXIMUM ERROR FOR IMPULSE TORQUE DISTURBANCES

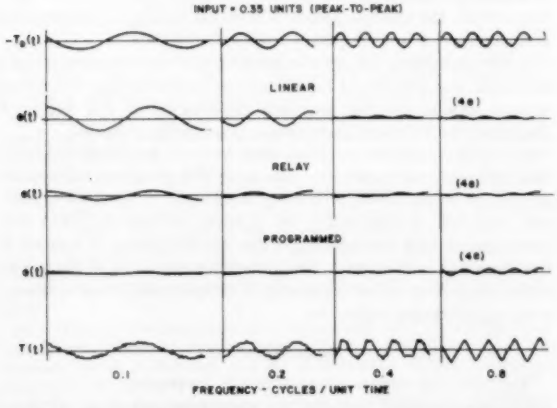


FIG. 19 RESPONSE OF THE THIRD-ORDER SYSTEM TO SINUSOIDAL TORQUE DISTURBANCES WITH LINEAR, RELAY, AND PROGRAMMED CONTROL

soidal torque disturbances whose rate of change does not exceed its rate limit. If the disturbance is expressed as

$$T_D = M \sin \omega t \quad [45]$$

then

$$\dot{T}_D = \omega M \cos \omega t \quad [46]$$

and a "critical frequency" ω_c can be defined as the frequency which makes the maximum value of $\dot{T}_D$ equal to the maximum possible rate of change of torque $\dot{T}_{max}$ or

$$\omega_c = \frac{\dot{T}_{max}}{M} \quad [47]$$

Observed critical frequencies are plotted for a number of disturbance magnitudes in Fig. 23, as well as a theoretical curve based on Equation [47]. Since the critical frequency is not easily identified experimentally, the agreement is perhaps satisfactory.

PARAMETER TOLERANCES

The programmed controller for a particular controlled system is unique. Given the dynamic behavior of the output member and torque source, there is only one switching surface which minimizes the response time for all step magnitudes. (The controller can be realized in a variety of ways by taking different combinations of e, $\dot{e}$, and $\bar{e}$ or related quantities as the inputs of the two-variable function generator.) This fact suggests that the load parameters must be known with great accuracy and that the switching surface must be constructed accurately. Although no systematic study of permissible errors has been carried out, a few pertinent remarks can be made.

The electro-optical function generator used in the analog-computer work was put into service without thoroughly checking its input-output relations against the theoretical switching surface; limited checks showed that the over-all error in a particular torque line, including the drafting errors in the disk, was sometimes as great as half the distance between lines. On the other

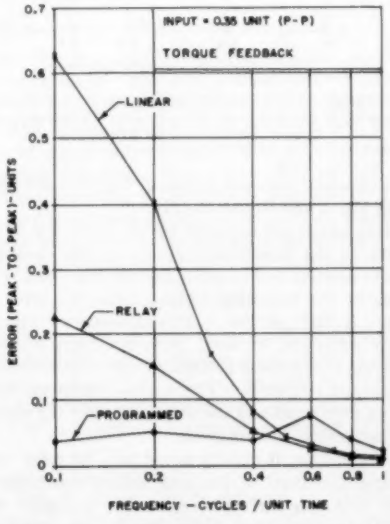


FIG. 20 RESPONSE OF THIRD-ORDER SYSTEM TO SINUSOIDAL TORQUE DISTURBANCES WITH LINEAR, RELAY, AND PROGRAMMED CONTROL USING TORQUE FEEDBACK

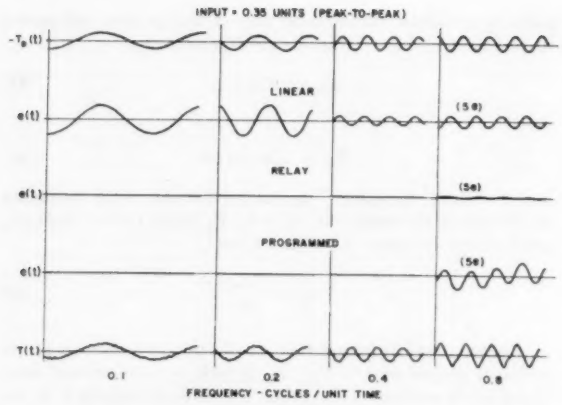


FIG. 21 RESPONSE OF THIRD-ORDER SYSTEM TO SINUSOIDAL TORQUE DISTURBANCES WITH LINEAR, RELAY, AND PROGRAMMED CONTROL USING ACCELERATION FEEDBACK

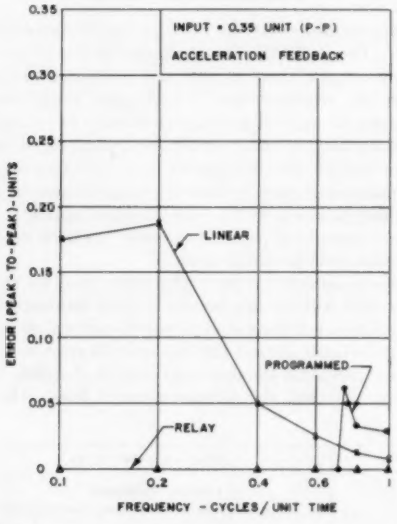


FIG. 22 RESPONSE OF THIRD-ORDER SYSTEM TO SINUSOIDAL TORQUE DISTURBANCES WITH LINEAR, RELAY, AND PROGRAMMED CONTROL USING ACCELERATION FEEDBACK

hand, the observed response curves of Fig. 11 are in good agreement with Fig. 3 and Equation [13].

In making minor adjustments on the computer, it was noted that drifts at the function-generator inputs caused different responses to positive and negative inputs; this effect is caused by translation of the switching surface along the error and error rate axes. A drift at the function generator output likewise shifted the switching surface. Misadjustment of an amplifier gain following the function generator resulted in either oscillatory or overdamped responses. The system remained stable when these errors were small and the deterioration of the step response, although noticeable, was not gre...

In this connection, it may be noted that the relay controller is a crude approximation to the programmed controller in which the proper switching surface is replaced by a plane. Misadjustments of the programmed system can be considered to distort the switching surface; provided that the distorted surface is intermediate between the two surfaces of Fig. 10, the response should likewise be intermediate between the programmed and relay systems.

In two experiments, a torque component proportional to the output velocity

$$T_f = \mathbf{f}\,\dot{c} \qquad [48]$$

was included; the viscous friction coefficient $\mathbf{f}$ was 0.1 and 0.5. With no change in the switching surface, only slight deterioration

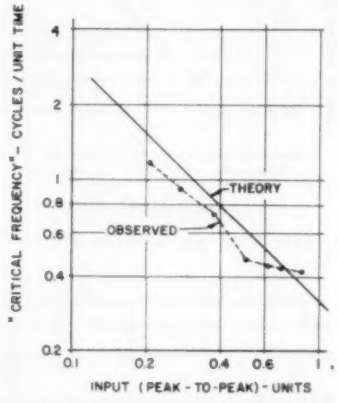


FIG. 23 CRITICAL FREQUENCY FOR RESPONSE OF PROGRAMMED THIRD-ORDER SYSTEM USING ACCELERATION FEEDBACK TO SINUSOIDAL TORQUE DISTURBANCES

of the step response was observed. The correct switching surface for an output member having inertia and friction was computed and found to be generally similar to the one used in these experiments, the region between $T = +1$ and $T = -1$ being narrower and closer to the error rate axis.

No experiments were conducted to determine the effect of output backlash, relay threshold, operating time or hysteresis, or other similar effects.

CONCLUSIONS

A programmed controller designed to minimize the response time of an idealized third-order system has been realized successfully for analog-computer investigations using an electro-optical two-variable function generator. The expected superiority of this type of control for step inputs was observed, and it was also found that the system behaved well for sinusoidal inputs and disturbances of low frequency and/or amplitude.

Some anomalous behavior, consisting of a type of subharmonic response, was discovered for sinusoidal inputs of large frequency and/or amplitude; the practical importance of this behavior could only be judged with reference to a specific situation.

A limited amount of direct and indirect evidence indicates that the required parameter tolerances in a programmed system are not as severe as was previously supposed. The effect of various parasitic nonlinearities on system performance was not investigated, and the ranking of the various modes of control is therefore not conclusive. Additional investigation of these and other matters would be necessary if programmed control theory were to be reduced to practice.

NOMENCLATURE

The following nomenclature is used in the paper:

c = controlled variable, or system output (e.g., airplane heading)

$\mathbf{c}$ = nondimensional system output

e = error
f = viscous friction coefficient
f = function of
J = moment of inertia
K = gain constant
K_a = acceleration gain constant
K_r = velocity gain constant
M = peak amplitude of periodic torque disturbance
R = magnitude of input step
s = Laplace operator
T = torque applied to the controlled system
T_c = computed torque
T_D = disturbance torque applied to system
T_m = maximum value of torque
t = time
$\mathbf{t}$ = nondimensional time
t_r = system response time
t_s = time at which last switching occurs
x, y, z = phase-space co-ordinates
ϵ = difference between computed and applied torque
τ = time required for torque to go from one extreme value to the other
τ_f = time constant (e.g., field time-constant of a d-c servomotor)
ω = angular frequency

BIBLIOGRAPHY

1 "Automatic Control System for Vehicles," by H. G. Doll, U. S. Patent 2,463,362.

2 "Textbook of Servomechanisms," by J. C. West, English Universities Press Ltd., London, England, 1953, p. 219.

3 "Nonlinear Techniques for Improving Servo Performance," by D. McDonald, Proceedings NEC, vol. 6, 1950, pp. 400–421.

4 "A Phase-Plane Approach to the Compensation of Saturating Servomechanisms," by A. M. Hopkin, Trans. AIEE, vol. 70, 1951, pp. 631–639.

5 "The Stabilization of On-Off Controlled Servomechanisms," by A. M. Uttley and P. M. Hammond, "Automatic and Manual Control," Academic Press, New York, N. Y., 1952, pp. 285–307.

6 "A Topological and Analog Computer Study of Certain Servomechanisms Employing Nonlinear Electronic Components," by R. C. Lathrop, PhD thesis, University of Wisconsin, 1951.

7 British Patent 29011, J. C. West, 1951.

8 Report No. 469, by D. W. Bushaw, Experimental Towing Tank, Stevens Institute of Technology, Hoboken, N. J., 1953.

9 "Optimization of Nonlinear Control Systems of Means of Nonlinear Feedback," by R. S. Neiswander and R. H. MacNeal, Trans. AIEE, vol. 72, part II, 1953, pp. 262–272.

10 "An Investigation of the Switching Criteria for Higher Order Contactor Servomechanisms," by I. Bogner and L. F. Kazda, Trans. AIEE, vol. 73, part II, 1954, p. 118.

11 "Nonlinear Optimization of Relay Servomechanisms," by L. M. Silva, Technical Report, Electronics Research Laboratory, University of California, Berkeley, Calif., April 15, 1954.

12 "Nonlinear Control of a Saturating 3rd-Order Servomechanism," by J. L. Preston, Technical Memorandum No. 6897-TM-14, Contract NOrd 11799, Servomechanisms Laboratory, M.I.T., Cambridge, Mass., May 24, 1954.

13 Discussion of Reference 10, R. Oldenburger.

14 "Effects of Friction in an Optimum Relay Servomechanism," by T. M. Stout, Trans. AIEE, vol. 72, part II, 1953, pp. 329–336.

15 "Graphical Solution of Some Automatic Control Problems Involving Saturation Effects With Application to Yaw Dampers for Aircraft," by W. H. Phillips, NACA Technical Note 3034, October, 1953.

16 "Use of Nonlinearities to Compensate for the Effects of a Rate-Limited Servo on the Response of an Automatically Controlled Aircraft," by S. F. Schmidt and W. C. Triplett. NACA Technical Note 3387, January, 1955.

17 "Sinusoidal Techniques Applied to Nonlinear Feedback Systems," by E. C. Johnson, Symposium on Nonlinear Circuit Analysis, Polytechnic Institute of Brooklyn, Brooklyn, N. Y., 1953; pp. 258–273.

18 "Analysis and Design Principles of Second and Higher-Order Saturating Servomechanisms," R. E. Kalman, AIEE Technical Paper 55–551, 1955.

19 "Synthesis of 'Optimum' Transient Response: Criteria and Standard Forms," by D. Graham and R. C. Lathrop, Trans. AIEE, vol. 72, part II, 1953, pp. 273–286.

20 "Introduction to Nonlinear Mechanics," by N. Minorsky, Edwards Brothers, Ann Arbor, Mich., 1947.

21 "Nonlinear Vibrations," by J. J. Stoker, Interscience Publishers, New York, N. Y., 1950.

22 "A Differential-Analyzer Study of Certain Nonlinearly Damped Servomechanisms," by R. R. Caldwell and V. C. Rideout, Trans. AIEE, vol. 72, part II, 1953, pp. 165–170.

23 "The Mechanism of Subharmonic Generation in a Feedback System," by J. C. West and J. L. Douce, Proceedings of the Institute of Electrical Engineers, vol. 102, 1955, part B, p. 569.

24 "Network Analysis and Feedback Amplifier Design," by H. W. Bode, D. Van Nostrand Company, Inc., New York, N. Y., 1945.

25 "A Study of a Predictor-Type Air-Frame Controller Designed by Phase-Space Analysis," by A. M. Hopkin and M. Iwama. AIEE Paper 56–198.

Discussion

RUFUS OLDENBURGER.[6] The Society is fortunate to have a paper by the author, who appears to be the first person to use switching functions for optimum transient response, and the co-author, who has contributed much to the science of the field. The work of these men, as well as that of other scientists in the servomechanism field, has been of inestimable value to both servomechanism and regulator engineers. Although the servomechanism specialist focuses his attention to a large extent on the problem of making an output follow an input, whereas the regulator expert concentrates on keeping the controlled quantity constant or nearly constant in value, the servomechanism specialist must concern himself with load disturbances, and the regulator expert with changes in the set point of the controller. The automatic control field owes much to the pioneering effort of the servomechanism scientist who has often led the way.

The optimum transients found by the writer in July, 1944, when he began working on the approach of the authors' paper, were for the system

$$Jc' + \alpha c = \beta m \quad \ldots\ldots [49]$$

where J, α, and β are constants and

$$| m' | \leq k_1 \quad \ldots\ldots [50]$$

for a constant k_1. Here c is the rpm of an airplane engine and m is propeller-blade angle. The primes denote derivatives with respect to time. The relation [50] means that the propeller servo speed is limited. Equation [49] is a torque equation, where J is the moment of inertia of the propeller shaft and connected parts and αc is a damping term. The system of Equations [49] and [50] is identical with that of the authors described by Equation [1], except for the damping term given by αc; i.e., Equation [49] corresponds to what the servomechanism experts term a "second-order" system, and which would then be written as

$$Jc'' + \alpha c' = \beta m' \quad \ldots\ldots [51]$$

where m' is bounded. The writer showed that the best transients are obtained by letting the propeller servo travel at maximum speed or zero speed at all times; i.e., we have m' equal to $\pm k_1$ or zero. In the normal servomechanism equivalent m' would be the torque T which takes on a maximum, minimum, or zero value at all times.

The switching surface given by Equation [28] of the paper for the system with Equation [23] is obtained by eliminating the variables t and t_s from Equations [27]. In order to reach equi-

[6] Director of Research, Woodward Governor Company, Rockford, Ill. Mem. ASME.

librium along a trajectory as described by Equations [25] and [27] it is necessary that one be on the switching surface in phase space. However, to reach equilibrium in an optimum manner, one must follow the correct curves on this surface, as traced back from the equilibrium point through Equations [25] and [27]. Thus if one has the initial conditions

$$e = e'' = 0, \qquad e' = e_0' > 0$$

and one assumes that one has a transient where $|T| < 1$, then equilibrium is attained with two switchings involving the functions

$$\Sigma_1 = e - e''\left(e' - \frac{e''^2}{3}\right) - \left(\frac{e''^2}{2} - e'\right)^{3/2}$$

$$\Sigma_2 = e' + \frac{\{e''\}}{2}$$

for the absquare $\{e''\}$ of e''; i.e.

$$\{e''\} = |e''|e''$$

The function Σ_1 comes directly from Equation [28], whereas Σ_2 is obtained by eliminating t, from the first and second Equations [26]. To obtain the optimum transient we use $e''' = -1$ for the first phase (see Fig. 24 of this discussion).

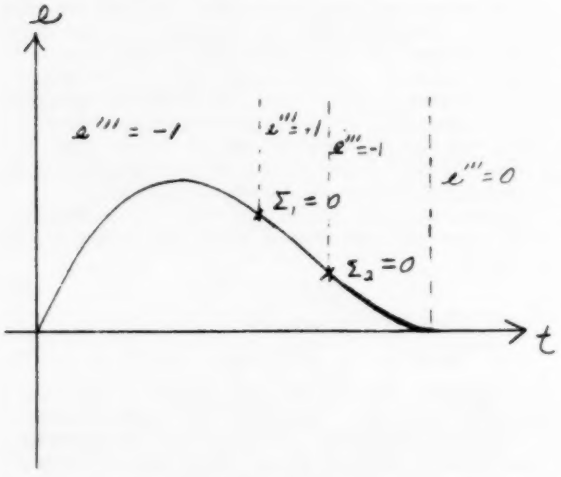


FIG. 24

When Σ_1 becomes zero we switch to $e''' = +1$. Finally, when we reach $\Sigma_2 = 0$, we let $e''' = -1$ until equilibrium with $e = e' = e'' = 0$ is reached. Then we switch to $e''' = 0$. The use of two switching functions is implied in the paper; i.e., the first switch puts one on the switching surface, and the second switch brings one from the Trajectory [27] to Trajectory [25]. However, once one is on the switching surface one is automatically led by the authors' method along an optimum trajectory to equilibrium. The writer and the authors are in complete agreement but are looking at the same thing from slightly different points of view.

If we introduce

$$\Sigma_3 = e - \frac{(e'')^3}{6}$$

obtained by eliminating t from the first and third Equations [25], then $\Sigma_2 = \Sigma_3 = 0$ implies that $\Sigma_1 = 0$. The Trajectory [25] is one for which $\Sigma_2 = \Sigma_3 = 0$.

The case just discussed is a limiting one for the single-lag systems (called "third order" by the authors) in the writer's ASME paper 56—IRD-13 entitled "Optimum Nonlinear Control." The writer has not carried through the work, but thinks that control can here be based on the sign of one function as he has proved holds for single-lag systems.

V. C. RIDEOUT[7] AND M. G. SPOONER.[7] The authors are to be congratulated on this thorough study of the third-order nonlinear servomechanism, and on their able combination of analytical and computer methods of attack upon this important problem.

The design of feedback systems in which nonlinearities are purposely exploited is still in its infancy, however. One evidence of this is the absence of compact criteria for comparison of such systems with one another and with linear systems. The authors, like other investigators[8, 9] have emphasized response to unit step-functions, although some attention also has been given to other wave forms and to rather vaguely defined random imputs. Their results bear out the well-known fact that nonlinear-system response to one type of input (or disturbance) may give no clue as to the response to another type of input.

In the rather common cases where inputs are statistically defined, and cannot be expressed in terms of semi-isolated step-functions or quasi-steady sinusoids, it would appear that correlation functions, which yield information as to mean-square error and delay, might serve more succinctly to state the degree of advantage of torque-switching servos over the corresponding linear servo. Some efforts[10] along these lines have been very useful in determination of equivalent linear systems, and at Wisconsin we have attempted[11] to show how this approach might be used to yield compact data on nonlinear-system performance. Further work along these lines is being pursued with the aid of a high-speed analog correlator.[12] It would be interesting to know if the authors have considered the use of such methods of study of their third-order nonlinear systems.

J. C. WEST.[13] The authors are to be congratulated on making this bold attempt to utilize a nonlinear function generator which is a function of two variables in a closed loop system.

There are two factors of great importance:

1 The operational results are not sensitive to deviations of the practical two variable function generator from the theoretical values.

2 Although based on step or transient-response optimization there is improved performance in other respects, i.e., for stochastic inputs and for sinusoidal signals of varying frequency.

It is interesting also as a confirmation of some theoretical work

---

[7] College of Engineering, University of Wisconsin, Madison, Wis.

[8] "Research in Nonlinear Mechanics as Applied to Servomechanisms," by P. E. Kendall and Irving Bogner, WADC Technical Report 53-521.

[9] "A Differential-Analyzer Study of Certain Nonlinearity Damped Servomechanism," by R. R. Caldwell and V. C. Rideout, Trans. AIEE, vol. 72, part II, 1953, pp. 165–170.

[10] "Nonlinear Servomechanisms With Random Inputs," by R. C. Booton, Jr., DACL Report No. 70, Massachusetts Institute of Technology, Cambridge, Mass., August 20, 1953.

[11] "The Use of Correlation Techniques in the Study of Servomechanisms," by T. M. Burford, V. C. Rideout, and D. S. Sather, *Journal of the British Institution of Radio Engineers*, vol. 15, May, 1955, pp. 249–257.

[12] "A High-Speed Correlator," by H. Bell and V. C. Rideout, Transactions of the Institute of Radio Engineers, (Electronic Computer Group), EC-3, No. 2, June, 1954, p. 30.

[13] Professor, Department of Electrical Engineering, Manchester University, Manchester, England.

by Bogner and Kazda[14] which shows that for an nth order differential equation the system must take $n-1$ reversals to bring error and derivatives of error to zero in the smallest time.

Work with a similar aim but approached in a different manner has been carried out by Coales and Noton[15] who have used prediction and a fast analog simulator built into the system in order to determine correct switching times.

The anomalous behavior of the system at high frequencies is of interest since the writer has made a study of these subharmonic phenomena. It has been found in general that for systems where the third subharmonic can be obtained it is also possible to obtain higher-order subharmonics[16, 17]—up to the 11th—by suitably increasing the input-driving frequency and increasing the amplitude. The range of amplitude and frequency over which the phenomenon is exhibited gets progressively smaller the higher the order of the subharmonic.

The writer should also like to ask if the authors found any "jump" phenomena in obtaining the fundamental frequency response, and if any anomalous behavior was observed under random testing conditions.

Authors' Closure

It is gratifying to the authors to find that Dr. Oldenburger and Dr. West are in agreement with them on the number of switches required to execute an optimum transient in a third-order system. A fairly general proof for the number of switching operations required, in agreement with previous assertions, has recently been given in a paper by Bellman, et. al.[18]

The authors, in their approach to this study, were initially most concerned with the criterion of response time following a step change of input or torque, since the design philosophy in this mode of nonlinear control is explicitly such as to minimize this response time. The authors did not make use of correlation functions to characterize the system performance; within the limited scope of the computer study reported on here, which was undertaken without reference to many of the considerations affecting the design and performance of a real system, the use of such sophisticated criteria as correlation functions did not appear to be justified. Furthermore, this study encourages the opinion that optimization of a nonlinear control with respect to one criterion does not necessarily mean that the control is "optimum" with respect to all other criteria.

Dr. West appears to have made a more thorough study of subharmonic phenomena than did the authors; jump phenomena were not observed when the fundamental frequency response was measured. The "anomalous" behavior observed under quasi-random conditions is illustrated in Fig. 15: As with sinusoidal inputs, when the rate of change of input exceeds the capacity of the torque source, anomalous behavior may be observed. The term anomalous merely implies that the theory of these phenomena is not completely understood.

Now that nonlinear control systems of this type have been examined in some detail and their principal characteristics are better understood, the authors hope and expect that their sphere of practical application will develop rapidly.

---

[14] Reference (10) of the paper.

[15] "An On-Off Servomechanism With Predicted Change-Over," by J. F. Coales and A. R. M. Noton, Proceedings of the Institute of Electrical Engineers, 1956, Paper No. M.1895.

[16] Reference (23) of the paper.

[17] "The Dual Input Describing Function and Its Use in the Analysis of Non-Linear Feedback Systems," by J. C. West, J. L. Douce, and R. K. Livesley, Proceedings of the Institute of Electrical Engineers, vol. 103, 1956, part B, Paper No. M.1877.

[18] "On the 'Bang-Bang' Control Problem," by R. Bellman, I. Glicksberg, and O. Gross, *Quarterly of Applied Mathematics*, vol. 14, April, 1956, pp. 11–18.

Optimum Nonlinear Control

By RUFUS OLDENBURGER,[1] ROCKFORD, ILL.

This paper is concerned with the response of a controlled system after an initiating disturbance has died out. Such a transient is obtained, for example, when the load on a prime mover is suddenly rejected or the speed setting of an engine governor is instantly switched to a new value. It is assumed that the rate of change of the controlling variable with respect to time is bounded, and that the maximum rate of change can be obtained arbitrarily. Thus the speed of a hydraulic governor servo is limited. The best return to equilibrium (minimum over or underswing, minimum duration of the transient, and so on) can be obtained under rather general conditions by having the servo or its equivalent travel only at maximum or zero speed. Control functions exist which give the optimum transients. These functions are nonlinear. The results of theoretical studies to enable the control designer to obtain optimum or nearly optimum transients are given here along with practical compromises. All results have been verified in the laboratory with physical devices (governors) of various kinds and automatically controlled systems.

Introduction

A CONTROL system must be stable. It is natural that much concern in the past has been with the problem of stability.* When a controller gives unstable performance the customer will simply not buy it. The use of derivative played an enormous role in the solution of the problem of stability. This problem is normally a linear one. Recently attention has been focusing on the problem of the quality of the control. Substantial improvement can be obtained in many areas by the use of nonlinear control.

To designate the class of controlled systems treated in this paper, the author will sometimes restrict himself to speed-governed prime movers. The results hold equally well for the control of temperature and other variables.

Consider a physical system with a single controlled variable. We consider the response of this system to a given disturbance, where after the disturbance dies out the system attains an equilibrium state S. We let m denote the deviation in the value of the controlled variable from its value in the state S. Similarly, let c be the deviation in the value of the controlling variable from the value it has for the state S. For the state S we then have $m = c = 0$. Let $P(D)$ and $Q(D)$ be polynomials in the derivative D with respect to time where the constant terms are different from zero. Let $e^{-\tau_d D}$ denote the dead-time operator, and let $g(c)$ be a monotonic nondecreasing function of c [this means that as c increases, the function $g(c)$ does not decrease]. By definition, given a function $c(t)$ of time t

$$e^{-\tau_d D}\, c(t) = c(t - \tau_d)$$

Let $f(m)$ be a monotonic nondecreasing function of m, and let the prime on m' denote the derivative dm/dt of m with respect to t. For systems in a single controlled variable the nonlinear theory developed by the author is devoted to those with the differential equation

$$m' + f(m) = \frac{P(D)e^{-\tau_d D}}{Q(D)}\, g(c) - L \qquad [1]$$

where L is a given (forcing) function of time. The polynomial $P(D)$ corresponds to leads and the polynomial $Q(D)$ to lags. Equation [1] may be considered to be a torque equation where m', $f(m)$, $[P(D)e^{-\tau_d D}/Q(D)]g(c)$, and L correspond to the inertial, damping, driving, and load torques, respectively, m being the revolutions per minute (rpm) of the prime mover and c the servo (throttle or equivalent) position. The coefficients in $Q(D)$ are assumed to be positive, since otherwise Equation [1] represents an unstable system. All of the applications encountered by the author in the speed-governor field are covered, at least to a first approximation, by Equation [1]. In most of the theory relating to Equation [1] the polynomial $P(D)$ is assumed to be a nonzero constant.

We suppose that the controlled system is at a given instant in a nonequilibrium state, and that the system thereafter reaches equilibrium while L is constant. It is assumed that c' is bounded. In the hydraulic speed-governor field this means the servo speed is bounded, which is always the case. It is also assumed that c' can arbitrarily be made to attain its maximum or minimum value, which is true for practical purposes in the case of hydraulic servos. Under rather general conditions the best transients in every sense (minimum over or underswing, minimum duration, etc.), i.e., *optimum control*, can be obtained by having c' take on only its maximum and minimum values and zero. In the case of speed governors this means that the servo should be permitted to travel at full speed or zero only.

Optimum control for systems with Equation [1] is not attained by having the servo travel at full or zero speed in the case where $P(D)$ is different from a constant. However, if the linear term in $P(D)$ is missing, satisfactory, or nearly optimum transients, may often be attained by having the servo travel at full or zero speed. When water hammer is accurately taken into account in the control of hydroturbines the polynomial $P(D)$ contains a linear term. Optimum transients are also not attained for a class of initial conditions unlikely to occur in practice, where lags are involved and m is approaching its equilibrium value at a numerically large rate m'.

Control functions depending on m and on the results of mathematical operations performed on m can be employed to yield optimum transients. These control functions are nonlinear, and are used to determine the values assigned to c' during a transient. Whether or not nonlinear control will be employed in specific cases depends on performance, engineering, and economic considerations.

In July, 1944, the author wished to determine how far the Woodward governors in use on the Hamilton Standard airplane propellers deviated from ideal ones. He immediately found that the best transients could be obtained by making the servo travel at full speed at all times. That is, for a sudden throttle burst or other instantaneous disturbance the servo should travel at full speed in one direction, then change at the correct instant

---

[1] Director of Research, Woodward Governor Company. Mem. ASME.

Contributed by the Instruments and Regulators Division of The American Society of Mechanical Engineers and presented at the ASME-AIEE Conference on Nonlinear Control Systems, Princeton, N. J., March 26–28, 1956.

Note: Statements and opinions advanced in papers are to be understood as individual expressions of their authors and not those of the Society. Manuscript received at ASME Headquarters, January 12, 1956. Paper No. 56—IRD-13.

to full speed in the other direction, and so on. The author obtained the ideal transients and established that the performance of the governors then in use was not too far from ideal, and that they were operating very satisfactorily.

Later the author developed the general theory of control where during transients the servo travels at full speed a maximum amount of time, the direction of motion being controlled by functions chosen to give optimum performance. The theory was checked in the laboratory in all of its details by the construction of controllers operating according to the desired functions.

We shall introduce the term *absquare* of a number x to mean the quantity $|x|x$. Here $|x|$ designates the absolute value of x. We denote the absquare of x by $\{x\}$. Thus the absquare of x is the signed square of x; i.e., $\{x\} = x^2$ when x is positive, and $-x^2$ when x is negative. For the simplest case of a controlled system, with no lags, the control function for optimum performance involves the absquare of the derivative m' of the deviation m of the controlled quantity. When the controlled quantity is prime-mover rpm, the quantity m' is prime-mover acceleration. A portion of a transient during which c' is constant, is termed a *phase*. For every system covered by Equation [1] with $P(D)$ equal to a constant, no dead time ($\tau_d = 0$), and $f(m)$ neglected [neglecting $f(m)$ is justified as theory shows], the absquare also occurs in the control function whose vanishing determines that c' should change sign at the start of the last phase before m reaches equilibrium (see Appendix 2). The same is true when the dead time is included and large load rejections are treated (Appendix 1). Because of these and other considerations it appears to the author that the absquare of the derivative m' may be the next element to be often, if not generally, added to control functions.

In so far as we know the first automatic control was the Watt governor. This control was proportional, then integral and derivative were added to controllers, and it now appears that the absquare may be employed. Finding that $f(m)$ could be neglected was quite important, since neglecting $f(m)$ made an enormous simplification in the theory.

In 1950 Donald McDonald (1)[2] published the control function which yields optimum transients in the simplest case of a system with no lags or leads. In this paper he used a heuristic phase-plane argument. In 1952 he published another paper (2) on the subject. This was a phase-plane study of torque saturation. In 1951 there appeared a paper by A. M. Hopkin (3) on the results of an empirical phase-plane analysis of a servo with limiting or saturation. A computer study of the subject was made by Richard C. Lathrop as part of his PhD thesis (4) the same year. At the Cranfield Conference in England in 1951, A. M. Uttley and P. H. Hammond (5) treated the switching of an on-off servo, employing the absquare. In a PhD thesis in 1953 Donald Wayne Bushaw (6) made a topological study of transients for the case of torque saturation controlled by switching functions, in particular, linear ones. Lawrence M. Silva studied switching functions to enable the servo to travel at full speed, and used an energy approach to find these functions (7 to 9). Irving Bogner and Louis F. Kazda (10) investigated the switching criteria for higher-order servomechanisms using the phase-plane approach, and showed that in general the number of switchings for optimum transients increases by one as the order of the controlled system goes up by one. Further work on higher-order servomechanisms was done by S. S. L. Chang (12) and R. E. Kalman (13). Servomechanisms with friction were treated by Louis F. Kazda (14) and T. M. Stout (15).

All of these contributions touch on the nonlinear theory developed by the author which for proprietary reasons has been kept confidential. The mathematics of this theory is far beyond the limits of this paper. However, for the benefit of the reader, the complete fundamental theory is given for the case of the simplest controlled system. It is proved that a best transient exists for any set of initial conditions, and that a control function Σ exists, such that m is brought to equilibrium along the best transient by having c' take on its minimum value when Σ is negative, its maximum when Σ is positive, and zero when Σ is zero. It is hoped that this treatment will give the reader confidence in the approach, and indicate the questions that must be answered to establish the theory for more complicated systems. Without a rigorous theory one cannot be sure that one has the right control functions and that a competitor will not do better. For more complicated systems the mathematical results are given without proof. The author feels that from these considerations the reader can obtain a fairly complete picture of why and where the nonlinear approach improves the quality of the control, and where the use of nonlinear terms is justified. Linear approximations are often satisfactory. The reader also should learn what compromises of the theory should be made in practice to preserve performance but reduce costs.

Controls in present use involve nonlinearities. Thus in many controllers c' takes on its maximum or minimum value when one is outside of a "control band" based on the value of m. Instead of these "unintentional" nonlinearities the correct ones for desired optimum performance should be built into the control.

It appears to the author that the development of new controls will proceed as follows: The correct control functions for optimum performance will be established by mathematical theory. The inherent complexity of this theory indicates that this will require the services of expert research mathematicians. After the correct control functions or reasonable compromises have been established the development engineer will incorporate them into practical devices.

Simple Control Problem

We shall first develop the theory for the simplest case and then proceed to more complicated problems. The simplest system is one where

$$m' = K_1 c \qquad [1]$$

for the controlled quantity m, controlling variable c, and constant K_1. As noted, m' is the derivative given by

$$m' = \frac{dm}{dt} \qquad [2]$$

In practice the rate of change c' of the controlling quantity is limited; i.e.

$$|c'| \leq K_2 \qquad [3]$$

for a constant K_2 and absolute value $|c'|$ of c'.

As an example we may have

m = engine rpm
c = throttle position

for a prime mover, where these quantities are deviations from equilibrium values. We use the term "throttle" in a general sense to denote rack position on a diesel engine, gate co-ordinate for a hydraulic turbine, fuel-valve position for a gas turbine, steam-valve position for a steam turbine, or throttle position for a gasoline engine. Equation [1] now says that the engine acceleration m' is proportional to throttle position (deviation from equilibrium). When the engine speed is subject to automatic control

[2] Numbers in parentheses refer to the Bibliography at the end of the paper.

the servo is connected to the throttle. It is therefore convenient to take

$$c = \text{servo position}$$

whence Equation [1] says that the engine acceleration is proportional to servo position.

Relation [3] says that the servo speed c' is bounded by K_2.

For example, in the case of the General Motors diesel GM71, driving a direct-connected alternator, with 1-in. servo movement from no-load to full-load, Equation [1] is (approximately)

$$m' = 600c \quad [4]$$

Thus if the servo is at the no-load position $c = 0$ and if the servo is moved suddenly to the full-load position $c = 1$, the engine accelerates at 600 rpm/sec. A typical Relation [3] for this case is

$$|c'| \leqq 10 \quad [5]$$

which means that the servo speed is limited to 10 ips; i.e., the servo will go from no-load to full-load in $^1/_{10}$ sec when traveling at maximum speed.

The condition that the servo speed is bounded is always true in practice. The power a servo can consume is limited and hence the servo speed is bounded. Sometimes the servo speed is limited by other considerations, such as to prevent excessive water hammer in hydro-power installations. In the case of aircraft-propeller governors, if the propeller changes pitch at too fast a rate, the passengers and pilot experience excessive discomfort. In the case of hydraulic speed governors the servo normally can be operated at full speed at will during a transient. This is also true, at least approximately, for other types of controls.

The problem of optimum transients is: Given any initial condition

$$t = 0,\ m = m_0,\ m' = m_0' \quad [6]$$

how must the servo c be moved so that equilibrium

$$m \equiv 0$$

will be reached in a minimum time with minimum overswing or underswing, and minimum area between the m-curve and the $m = 0$ axis? We shall show that for each set of initial conditions there is a unique transient with m-curve Γ such that *acceptable criteria of automatic control theory for optimum transients are satisfied simultaneously.*

Optimum Transients for Simple Systems

We shall prove later that for an optimum transient in the case of the System [1] and [3] the servo must travel at all times at full speed until equilibrium is attained. Hence for such a transient

$$c' = \pm K_2 \quad [7]$$

It will be no restriction on the generality of the method to assume that

$$m_0 \geqq 0$$

whence at the start of the transient one is above or on the t-axis. Let us assume that the m-curve is concave up before equilibrium so that we come into equilibrium from above the t-axis as shown in Fig. 1. The curve Γ leading to equilibrium is thus concave up before equilibrium. Then

$$c' = K_2 \quad [8]$$

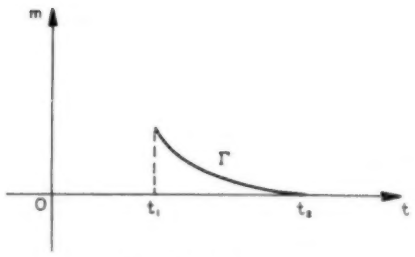


Fig. 1 Arc of Ideal Curve

and by Equation [1] the transient satisfies the equation

$$m'' = K_1K_2 \quad [9]$$

The second derivative m'', i.e.

$$\frac{d^2m}{dt^2}$$

is the rate of change of acceleration m' with respect to time t.

To solve Equation [9] it will be convenient to make some transformations in the variables t, c, and m. Otherwise the formulas will be excessively complicated. We can write Relations [1] and [3] as

$$\left(\frac{m}{K_1K_2}\right)' = \left(\frac{c}{K_2}\right), \qquad \left|\left(\frac{c}{K_2}\right)'\right| \leqq 1 \quad [10]$$

Introducing new variables M and C so that

$$M = \frac{m}{K_1K_2}, \quad C = \frac{c}{K_2} \quad [11]$$

the Relations [10] become

$$M' = C, \qquad |C'| \leqq 1 \quad [12]$$

which are easier to manipulate. We shall suppose that the curve Γ is concave down before the instant $t = t_1$ and that equilibrium is reached at the instant $t = t_2$. We introduce a new time scale T where

$$T = t - t_1$$

so that at the start of the arc of Fig. 1 we have $T = 0$. We note this does not change Relations [12] since

$$\frac{dM}{dt} = \frac{dM}{dT}, \qquad \frac{dC}{dt} = \frac{dC}{dT} \quad [13]$$

We shall therefore understand that the primes in Relations [12] denote derivatives with respect to T. Equation [9] now becomes

$$M'' = 1 \quad [14]$$

Let M_1 and M_1' denote the values of M and M' at $T = 0$. By the theory of differential equations (11) the solution of Equation [14] is

$$M' = M_1' + T \quad [15]$$

$$M = M_1 + M_1'T + \frac{T^2}{2} \quad [16]$$

When M' reaches zero we want M to reach zero simultaneously;

i.e., the minimum point of the arc in Fig. 1 should be on the t-axis. The three possibilities of solutions that emanate from

$$T = 0, \quad M = M_1 \ldots\ldots\ldots\ldots\ldots\ldots [17]$$

at the point P_1 are shown in Fig. 2.

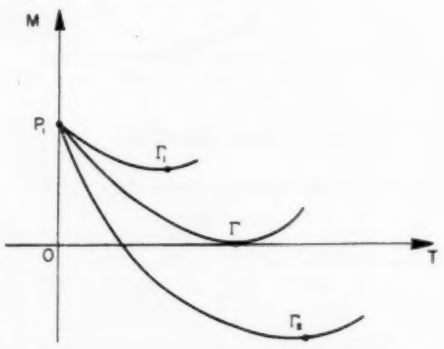


FIG. 2 ARCS FOR DIFFERENT INITIAL SLOPES

If we leave the point P_1 with too steep a slope, we obtain the curve Γ_2 (corresponding to the Equations [15] and [16]) which overshoots the T-axis giving an overswing accordingly. If the slope at P_1 is too little the corresponding curve Γ_1 undershoots the T-axis as shown. The slope M_1' must be chosen so that $M = 0$ at the same instant that $M' = 0$. By Equation [15] we have $M' = 0$ when

$$T = -M_1' \ldots\ldots\ldots\ldots\ldots\ldots [18]$$

Substituting T from Equation [18] in Equation [16] and imposing the condition that $M = 0$ when T satisfies Equation [18] we have

$$M_1 - \frac{(M_1')^2}{2} = 0 \ldots\ldots\ldots\ldots\ldots\ldots [19]$$

It follows that if M_1 and M_1' are not both zero and are related at any instant by the Equation [19] and if thereafter we keep

$$C' = 1 \ldots\ldots\ldots\ldots\ldots\ldots [20]$$

we shall reach $M = 0$ at the same instant that $M' = 0$. If after this instant we keep

$$C' \equiv 0 \ldots\ldots\ldots\ldots\ldots\ldots [21]$$

that is, we hold the servo still, the variable M will satisfy the relation

$$M'' \equiv 0 \ldots\ldots\ldots\ldots\ldots\ldots [22]$$

whence $$M \equiv 0$$

so that we remain at equilibrium until the system is disturbed.

If we come into equilibrium from below the T-axis and the servo is traveling in one direction at full speed from $T = 0$ to equilibrium, where equilibrium is reached for some positive value of T, we have

$$C' = -1 \ldots\ldots\ldots\ldots\ldots\ldots [23]$$

for this arc, see Fig. 3.

Replacing M_1 by $-M$ we obtain

$$M + \frac{(M')^2}{2} = 0 \ldots\ldots\ldots\ldots\ldots\ldots [24]$$

from the Condition [19].

Thus if $M' > 0$ we have Equation [24], and if $M' < 0$, Equation [19] with the subscripts on M_1 and M_1' omitted. These equations can therefore be combined into

$$M + \frac{1}{2}\{M'\} = 0 \ldots\ldots\ldots\ldots\ldots\ldots [25]$$

i.e.

$$M + \frac{1}{2}|M'|M' = 0$$

If at any instant M and M' satisfy Equation [25] the Relation [20] or [23] will lead to equilibrium, where Equation [20] is used if $M > 0$, and Equation [23] if $M < 0$.

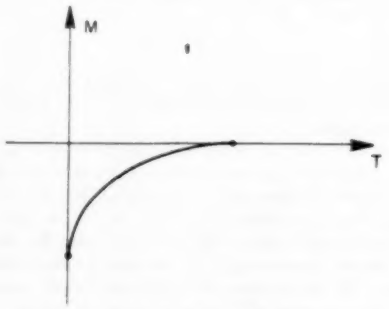


FIG. 3 CURVE CONCAVE DOWN

We introduce the notation Σ where

$$\Sigma = M + \frac{1}{2}\{M'\} \ldots\ldots\ldots\ldots\ldots\ldots [26]$$

The *derivative of the absquare* $\{x\}$ is $2|x|x'$. Differentiating Σ with respect to T we obtain

$$\Sigma' = M' \pm M'M'' = (1 \pm M'')M' \ldots\ldots\ldots [27]$$

where the $\pm$ sign is $+$ when $M' > 0$ and $-$ when $M' < 0$. For the arcs of Figs. 1 and 3 the sign of M'' is opposite the sign of M'. Thus for these arcs the $\pm$ sign in Equation [27] is opposite the sign of M''. But

$$M'' = \pm 1 \ldots\ldots\ldots\ldots\ldots\ldots [28]$$

Hence for the arcs of Figs. 1 and 3 we have

$$\Sigma' = 0 \ldots\ldots\ldots\ldots\ldots\ldots [29]$$

whence Σ is a constant. Since $\Sigma = 0$ when equilibrium is reached it follows that

$$\Sigma \equiv 0 \ldots\ldots\ldots\ldots\ldots\ldots [30]$$

for these arcs, whence Relation [25] holds everywhere along these arcs.

Suppose that we have a transient curve Γ as in Fig. 1, and that prior to $T = 0$ we have Equation [23] valid. The curve is then concave down before $T = 0$. Keeping Equation [23] valid we trace the curve Γ back to a point Q where $M = 0$ as shown in Fig. 4. There will always be such a point Q as the following argument shows. Since

$$M'' = -1 \ldots\ldots\ldots\ldots\ldots\ldots [31]$$

to the left of the M-axis, the slope of the curve Γ extended to the left increases by 1 for each unit of time T as we travel to the left.

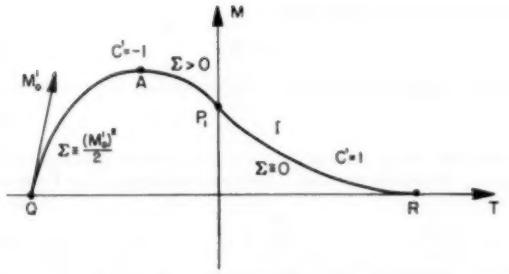


FIG. 4 IDEAL TRANSIENT ABOVE T-AXIS

The slope M' will thus never reach infinity, and eventually the extended curve Γ will cross the M-axis with a positive slope M_0'.

The high point on Γ extended is denoted by A. From Q to A we have $M' > 0$ whence

$$\Sigma' = (1 + M'')M' = 0 \quad \ldots\ldots [32]$$

so that Σ is a constant. Since at Q we have

$$M = 0, \qquad M' = M_0' > 0 \quad \ldots\ldots [33]$$

it follows that

$$\Sigma \equiv \frac{1}{2}(M_0')^2 > 0 \quad \ldots\ldots [34]$$

from Q to A. From A to P_1 we have $M' < 0$ and $M'' = -1$ whence

$$\Sigma' = (1 - M'')M' = 2M' < 0 \quad \ldots\ldots [35]$$

Thus Σ is decreasing from the value it has along the arc QA to the value zero, which it has along the arc P_1R.

The curve shown in Fig. 4 leads from the point Q to the equilibrium point R with one switching of the direction of motion of the servo. Suppose now that the system is disturbed so that at a given instant M and M' satisfy Relations [33] where M_0' is an arbitrary positive number. The curve of Fig. 4 (normally the switch point P_1 will not be on the M-axis) will automatically lead to equilibrium if we take

$$\left.\begin{aligned} C' &= -1, \quad \Sigma > 0 \\ C' &= +1, \quad \Sigma \equiv 0 \end{aligned}\right\} \quad \ldots\ldots [36]$$

Suppose now that at a given instant T_a we have

$$M = M_a > 0, \quad M' = M_a', \quad \Sigma = \Sigma_a > 0 \quad \ldots\ldots [37]$$

With $M'' = -1$ we can trace a curve to the left of $T = T_a$ of the form shown in Fig. 4, until we reach $M = 0$. Thus the element $(M, M') = (M_a, M_a')$ can be associated with a curve that starts at $M = 0$ as shown in Fig. 5. The slope M_a' at the point P_a can be either $+$, 0, or $-$ as long as Relations [37] are satisfied. It follows that we can go from P_a to an equilibrium point R by following a curve as shown in Fig. 4; i.e., by choosing the sign of C' according to Relations [36]. If at P_a the slope M_a' is such that $\Sigma = 0$, instead of > 0, we reach equilibrium along the curve for which $C' = 1$. We have thus taken care of all initial conditions where

$$M_a \geqq 0, \qquad \Sigma_a \geqq 0 \quad \ldots\ldots [38]$$

If

$$M_a \leqq 0, \qquad \Sigma_a \leqq 0 \quad \ldots\ldots [39]$$

we can take a curve, as shown in Fig. 6, through the point P_a that will lead to the equilibrium point R and when traced to the left originates at a point Q where

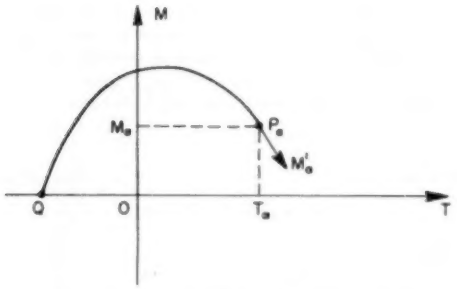


FIG. 5 CURVE TRACED FROM POINT P_a TO Q

$$M = 0, \qquad M' = M_0' < 0 \quad \ldots\ldots [40]$$

Suppose now that at the point P_a of Fig. 7 we have

$$M_a > 0, \qquad \Sigma_a < 0 \quad \ldots\ldots [41]$$

It follows that

$$M_a' < 0 \quad \ldots\ldots [42]$$

From P_a we trace a curve Γ'' to the right for which $C' = 1$. In the Derivative [27] for Σ' we have the minus sign and $M'' = 1$, whence Equation [29] holds. Since M' increases by one unit for each unit increase in T, eventually we will reach a point S where $M' = 0$. But then $M < 0$ since along Γ''

$$\Sigma \equiv \Sigma_a < 0 \quad \ldots\ldots [43]$$

It follows that $M = 0$ for some point Q between P_a and S. The

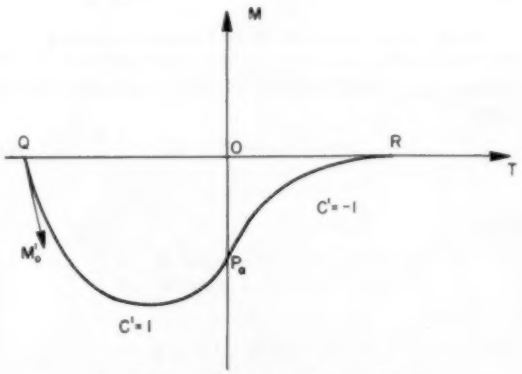


FIG. 6 IDEAL TRANSIENT BELOW T-AXIS

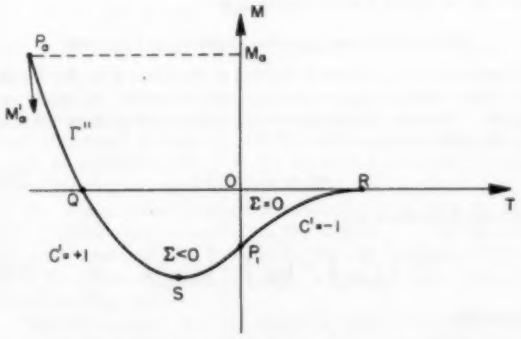


FIG. 7 IDEAL TRANSIENT WITH NEGATIVE OVERSHOOT

curve Γ'' is of the type Γ_2 of Fig. 2. We can now reach equilibrium as in Fig. 6. The entire transient is then as shown in Fig. 7, provided that the T-co-ordinate of the point P_o is chosen properly. For the curve of Fig. 7 $C' = 1$ from P_o to P_1 and $C' = -1$ from P_1 to R. For the arc P_oP_1 we have

$$C' = 1, \qquad \Sigma < 0 \ldots\ldots\ldots\ldots\ldots [44]$$

and for P_1R

$$C' = -1, \qquad \Sigma = 0 \ldots\ldots\ldots\ldots\ldots [45]$$

Similarly, if

$$M_o < 0, \qquad \Sigma_o > 0 \ldots\ldots\ldots\ldots\ldots [46]$$

we have a curve as shown in Fig. 8, provided that the T-co-ordinate of the point P_o is chosen so that the switch point P_1 is on the M-axis.

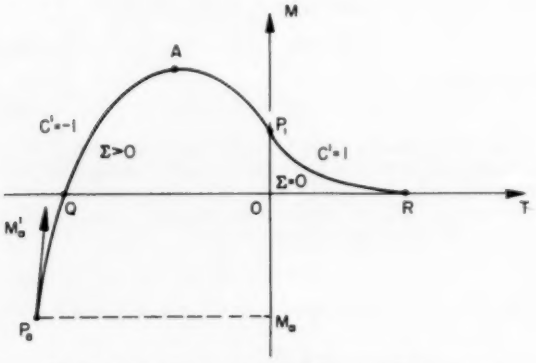


FIG. 8 IDEAL TRANSIENT WITH POSITIVE OVERSHOOT

We have now treated all initial conditions and shown that for all cases

$$C' = 1 \text{ when } \Sigma < 0 \ldots\ldots\ldots\ldots\ldots [47]$$

$$C' = -1 \text{ when } \Sigma > 0 \ldots\ldots\ldots\ldots\ldots [48]$$

whereas

$$C' = 1 \quad \text{when } \Sigma = 0 \text{ and } M > 0$$
$$C' = -1 \text{ when } \Sigma = 0 \text{ and } M < 0$$

and

$$C' = 0 \text{ when } \Sigma = M = 0$$

The ideal transients of Figs. 4 and 6 to 8 are determined entirely by the sign of Σ except when $\Sigma = 0$.

LOAD REJECTIONS AND SPEED-SETTING CHANGES

Transients of Figs. 4 and 6 starting at the point Q on the T-axis arise when instant load rejections and increases are made on engines. To take care of load we introduce a load term $-l$ into Equation [1] to obtain

$$m' = K_1c - l \ldots\ldots\ldots\ldots\ldots [49]$$

which can be written as

$$\left(\frac{m}{K_1K_2}\right)' = \left(\frac{c}{K_2}\right) - \left(\frac{l}{K_1K_2}\right) \ldots\ldots\ldots\ldots [50]$$

from which

$$M' = C - L \ldots\ldots\ldots\ldots\ldots [51]$$

when we make the substitution

$$l = K_1K_2L \ldots\ldots\ldots\ldots\ldots [52]$$

and the Transformations [11]. In equilibrium $M' = 0$ whence

$$C = L \ldots\ldots\ldots\ldots\ldots [53]$$

so that the servo takes a unique position corresponding to the load L. Dropping the load L is equivalent to replacing Equation [51] by

$$M' = C \ldots\ldots\ldots\ldots\ldots [54]$$

given in Relations [12]. Suppose that we are in equilibrium before load rejection so that

$$C = L$$

Dropping the load L instantly we have

$$M' = L \ldots\ldots\ldots\ldots\ldots [55]$$

from Relation [54] at the start of the transient. Hence at the initial point Q of the transient we have

$$M_0' = L \ldots\ldots\ldots\ldots\ldots [56]$$

Thus the curve of Fig. 4 is the response to an instant load rejection. Transients starting at a point off the T-axis may arise when one instant load rejection is followed by another before the response to the first rejection has died out.

Transients starting at a point off from the T-axis with $M' = 0$ arise when there is an instant speed-setting change. Such a transient is shown in Fig. 9 where the transient starts at $T = 0$, $M = -M_0$.

We shall prove in the next section that the curves of Figs. 4, and 6 to 9 are optimum curves.

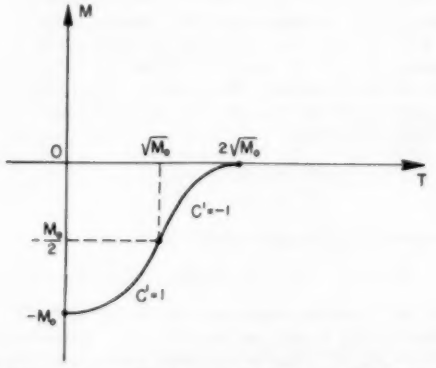


FIG. 9 RESPONSE TO CHANGE IN SETTING OF THE CONTROLLED VARIABLE

For the original variables and instant load rejections the results are shown in Fig. 10.

In a linear control the duration (properly defined) of a transient is independent of the magnitude of the disturbance, and the maximum deviation from equilibrium for instant load rejections is proportional to the magnitude of the disturbance, i.e., to m_0', as shown in Fig. 11. Thus if the maximum deviation from equilibrium for a given load change is m_M, and if the load change is halved, the deviation is also halved to $^1/_2m_M$, but the maximum still occurs at the same instant t_M.

For the nonlinear control of Fig. 10 if we halve the magnitude of the load rejection, *the deviation m_M from equilibrium goes to* $^1/_4m_M$, i.e., *as the square of the magnitude of the disturbance,* and

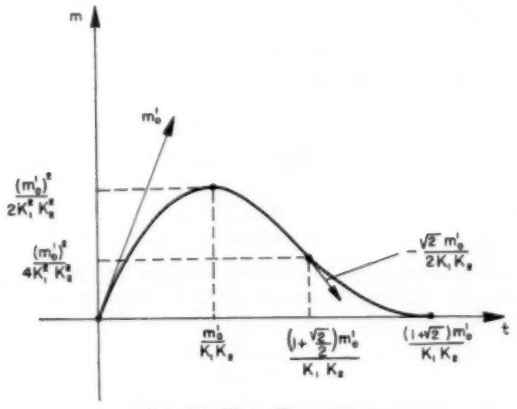


FIG. 10 IDEAL TRANSIENT

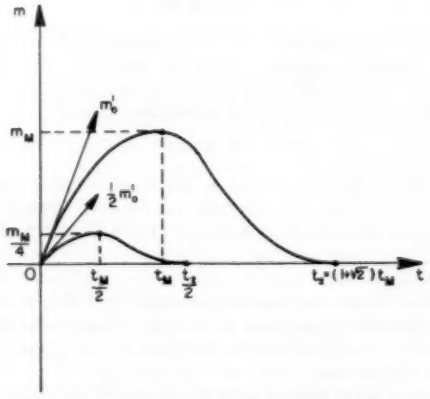


FIG. 12 TRANSIENTS FOR NONLINEAR CONTROL

the duration of the transient goes down by half, i.e., *in proportion to the magnitude of the disturbance*, as shown in Fig. 12.

For an instantaneous increase in the setting of the controlled variable the optimum curve is shown in Fig. 13.

WHY THE NONLINEAR TRANSIENTS ARE OPTIMUM

To prove that the "nonlinear" transients obtained in this paper are optimum, let us consider any state of the system, i.e., at any time which we may take to be $T = 0$, we have

$$M = M_0, \qquad M' = M_0' \ldots\ldots\ldots\ldots [57]$$

given. This means that the vector in Fig. 14 is given. In this figure we have drawn the "optimum" nonlinear transient Γ as derived in this paper. There is a unique such curve Γ leading to equilibrium at the point P. In Fig. 14 the curve is concave down first ($C' = -1$) up to the point P_1 with $T = T_1$, and then concave up ($C' = +1$). Consider a trajectory Γ_1 that leaves the point $(0, M_0)$ with the slope M_0' but reaches equilibrium before the point P. The curve Γ_1 must then cross or leave the curve Γ at some point P_2 before the time $T = T_2$ associated with P. This point P_2 is to the right of the point P_1 where the concavity of Γ changes. This must be so because the slope cannot change faster along a trajectory than along Γ between the initial point P_0 and the point P_1. If we start at the point P_2 and increase the slope at the maximum rate we cannot reach $M' = 0$ before $T = T_2$. The curve Γ_1 will be below Γ at $T = T_2$ as shown in Fig. 15.

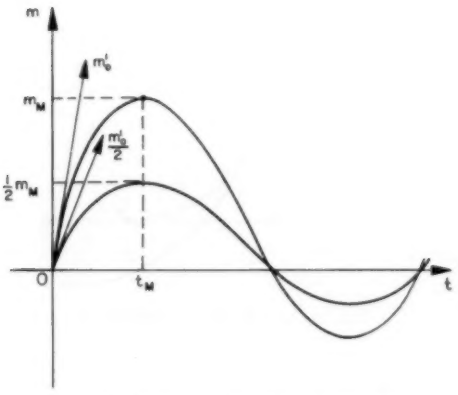


FIG. 11 TRANSIENTS FOR LINEAR CONTROL

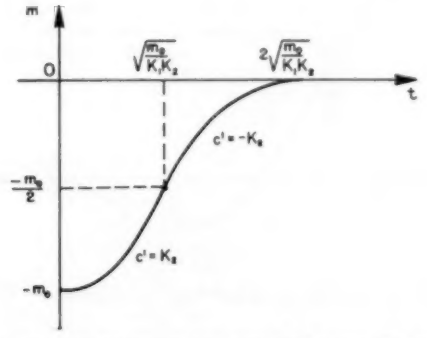


FIG. 13 OPTIMUM TRANSIENT FOR INSTANT CHANGE IN SETTING OF CONTROLLED VARIABLE

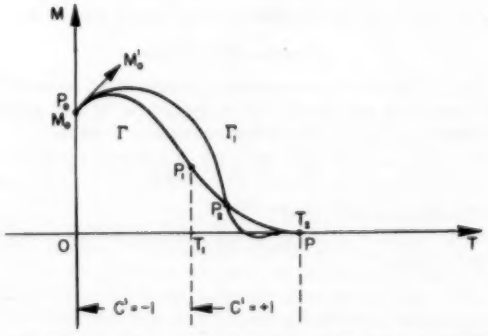


FIG. 14 TRAJECTORIES Γ AND Γ_1 THROUGH GIVEN ELEMENT (M_0, M_0')

Suppose that the curve Γ undershoots as shown in Fig. 16. Any trajectory leaving P_0 will be below or pass through the point P_1 at the instant $T = T_1$, where the concavity changes, since the slope along Γ is increasing at the maximum rate up to P_1. For a curve Γ_1 starting at the point P_0 with the slope M_0' to reach equilibrium before the instant $T = T_2$ the curve Γ_1 must cross or leave Γ between P_1 and the point P where $T = T_2$. This cannot be so in view of the argument relative to the case of Fig. 14.

The best transient in every case is thus one where equilibrium is reached in one phase with $C' = 1$ or $C' = -1$, or in two phases

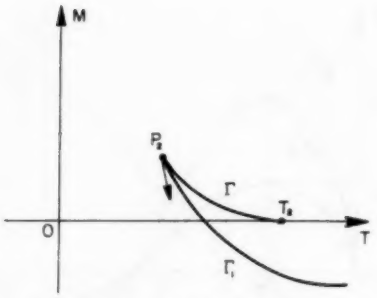


FIG. 15 ARC Γ_1 WITH UNDERSHOOT

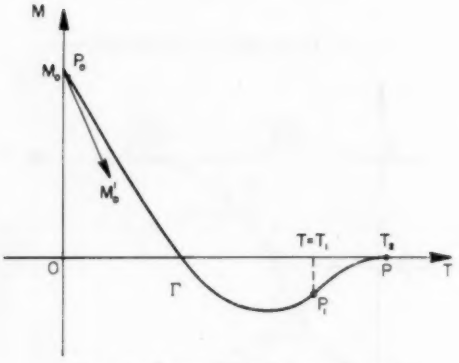


FIG. 16 CURVE Γ WITH UNDERSHOOT

with a phase $C' = 1$ followed by a phase for which $C' = -1$, or a phase with $C' = -1$ followed by a phase with $C' = +1$. Here, as in the introduction, phase refers to a portion of the transient for which C' is a constant (see Figs. 4, 6 to 10, and 13).

CONTROL FUNCTIONS

The problem is now to construct a control which senses M only, and quantities such as M' which depend on M and yields the optimum transients. To do this we consider Σ where

$$\Sigma = M + {}^1/_2\{M'\} \quad [58]$$

for the absquare $\{M'\}$ of M'.

Along the arc P_oA of Fig. 8 (from the initial point to the maximum) we have

$$\Sigma' \equiv 0 \quad [59]$$

whence

$$\Sigma \equiv M_o + {}^1/_2\{M_o'\} \quad [60]$$

Along the arc AP_1 from the maximum point A to the point of inflection P_1 we have

$$\Sigma' = 2M' \quad [61]$$

and from the point P_1 to the equilibrium point R we have

$$\Sigma' \equiv 0 \quad [62]$$

whence

$$\Sigma \equiv 0. \quad [63]$$

since at R we have

$$M = M' = 0 \quad [64]$$

We now use the control

$$C' = -K\Sigma \quad [65]$$

where K is "infinitely" large, i.e.

$$\left.\begin{aligned} C' &= -1 \text{ for } \Sigma > 0 \\ C' &= \ \ 0 \text{ for } \Sigma = 0 \\ C' &= +1 \text{ for } \Sigma < 0 \end{aligned}\right\} \quad [66]$$

There is an apparent contradiction between the Schedule [66] and $C' = +1$ along the arc P_1R of Fig. 8. However, in practice this is not the case. Consider the inflection point P_1 where we set

$$C' = 0 \quad [67]$$

in view of Schedule [66]. When Equation [67] holds we have

$$M'' = 0$$

whence M' is a constant. Since M is decreasing Σ will "immediately" become negative. Then by Schedule [66] we have $C' = +1$ and we follow the arc P_1R to equilibrium.

In any physical situation we would have at least

$$\Sigma = -\epsilon \quad [68]$$

for a small positive number ϵ before we would switch from $C' = -1$ to $C' = +1$. Since Σ is a constant along an arc for which $C' = +1$ and $M' < 0$, when we reach $M' = 0$ we have

$$\Sigma = M = -\epsilon \quad [69]$$

Thus the transient will undershoot as shown in Fig. 17 where we have taken the initial point at $T = 0$. The amount of undershoot will thus depend on the deadband in the device that controls the servo speed.

Unless a linear zone or equivalent is provided to stabilize the system an on-off servo will hunt forever after a sizable disturbance as can be readily verified by mathematical theory. In practical devices built by the author and his associates for the simple system under discussion, the schedule given by the Formulas [66] is used except that when M is small in absolute value the M-term in Σ is replaced for a constant a by aM and when $|M'|$ is small enough the absquare term is replaced by bM' for a constant b so that the function C' given by

$$C' = -K(aM + bM') \quad [70]$$

yields good characteristic roots when combined with

$$M'' = C' \quad [71]$$

For extremely good electronic components, and a system with no lag for practical purposes, one can use a value of K as high as 200, and still have stable transients for disturbances where $M = 0$, $M' = 1$. However, 200 is near the upper limit. See Fig. 18 for the case of optimum transients and instant load rejections obtained in the laboratory. Fig. 18 and Figs. 21 to 27 are photographs of Sanborn oscillograph traces. The RPM-curve of Fig. 18(a) shows the rpm response to instant 25 per cent load rejections and load increases, whereas the RPM-curve of Fig. 18(b) shows the corresponding results for the 50 per cent load case. Note that the RPM-overswing for 25 per cent load rejection is one fourth as much as for 50 per cent load rejection. Note also that there are no RPM-underswings for the load rejections, and that the transients for the 25 per cent load case are over in one half of the time required for the 50 per cent case. The servo

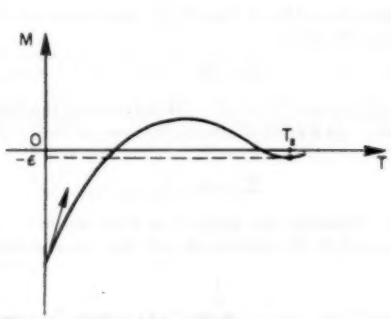


FIG. 17 PRACTICAL TRANSIENT

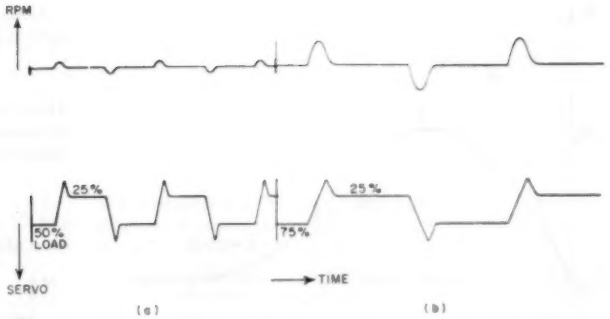


FIG. 18 PHOTOGRAPH OF OPTIMUM NONLINEAR TRANSIENTS IN AN ELECTRONIC CASE

trace of Fig. 18(*a*) is to a different vertical scale from this trace for Fig. 18(*b*). For Fig. 18(*a*) the servo goes from the 50 per cent load level to the 25 per cent load level, then back to the 50 per cent level, etc. The levels for Fig. 18(*b*) are for 75 and 25 per cent load.

The optimum physically realizable characteristic roots can be determined from theory and the limitations of the equipment. The change from M to aM in Σ can normally be made at M between 0.001 and 0.1, and similarly the change from $\frac{1}{2}\{M'\}$ to bM' can be made when M' is somewhere in the range from $M' = 0.01$ to $M' = 0.1$. The place where a change in a coefficient is made depends on the magnitude of the disturbances. For the case of Fig. 18 the lags in the system and control were very small so that without a linear control band with the Function [70] the hunt in M was 0.0025 per cent peak to peak. The range of load rejections was 0.625 to 50 per cent, i.e., a range of 80:1 so that the range in overspeeds was 6400:1, namely, from 0.0078 to 50 per cent.

Consider, for the moment, the case where ϵ is infinitely small (to be rigorous one should use a limit process here). Suppose that the system at any instant is in the state $M = M_0$, $M' = M_0'$, say, at $T = 0$. If $\Sigma > 0$, then by the Schedule [66] we have $C' = -1$. Along the trajectory thus obtained Σ will be constant if $M' \geq 0$. Eventually, $M' < 0$, whence Σ will decrease to $\Sigma = 0$. By the schedule we then have $C' = 0$, and Σ becomes negative immediately, whence by the Schedule [66] we soon have $C' = +1$ and we are brought to equilibrium. At the instant when equilibrium is reached we set $C' = 0$, whence $\Sigma = 0$ thereafter, at least until the next disturbance; similarly, if $\Sigma < 0$ at $T = 0$.

If $\Sigma = 0$ at $T = 0$ and we are not at equilibrium, then

$$M_0' \neq 0 \qquad [72]$$

and with $C' = 0$ immediately Σ becomes + or —, and Schedule [66] will lead to equilibrium along an optimum transient.

Since all transients are optimum when the Schedule [66] is used, we obtain "optimum" transients for instant load rejections as well as for instant changes in the setting of the controlled variable by using this schedule.

EFFECT OF DAMPING IN SYSTEM

The same type of mathematical treatment as used in the simplest case holds for other cases. Because this theory is extremely complicated and beyond the limits of this paper the proofs will be omitted although the results will be summarized.[3]

Consider now the system with the equation

$$M' + \alpha M = C \qquad [73]$$

with $\alpha \geq 0$ in place of the equation in Formulas [12]. The term αM in this equation is a damping term. We introduce the controlling function Σ where

$$\Sigma = M + \frac{M'}{\alpha} \pm \frac{\ln(1 + \alpha|M'|)}{\alpha^2} \qquad [74]$$

Here the $\pm$ sign is

$$\left.\begin{array}{l} \text{plus when } M' < 0 \\ \text{minus when } M' > 0 \\ \text{either when } M' = 0 \end{array}\right\} \qquad [75]$$

Expanding the ln-term of Equation [74] in powers of $\alpha|M'|$ we obtain

$$\Sigma = \Sigma_1 + \alpha\left(\frac{-\{M'\}|M'|}{3} + \frac{\alpha\{M'\}|M'|^2}{4} - \frac{\alpha^2\{M'\}|M'|^3}{5} \cdots\right) \qquad [76]$$

where

$$\Sigma_1 = M + \tfrac{1}{2}\{M'\} \qquad [77]$$

is the control function for the case with the damping term neglected. If $|M'|$ is small the function Σ can thus be replaced by Σ_1. In any case it is advisable to use the function Σ_1 of Formula [77] in place of the Σ of Formula [74] as if the damping term were missing. *Theory shows that the presence of the damping improves the transients and that the difference in neglecting and not neglecting the damping term is often not too great.* Similarly, if the damping term is $f(M')$ for a monotonic nondecreasing function of M' where

$$f(0) = 0$$

it is advisable to neglect the damping, since theory shows its presence to be beneficial.

Mathematical considerations show that the optimum transient for a load rejection is that given in Fig. 19. Along the arc AP_1 we have

$$\Sigma' = \frac{M'(2 - \alpha M')}{1 - \alpha M'} \qquad [78]$$

whence Σ is decreasing. At the point P_1 (not in general an inflection point) Σ becomes zero. Schedule [66] still applies, however, with the Σ of Formula [74].

[3] The proofs are on hand but have not been written up for publication.

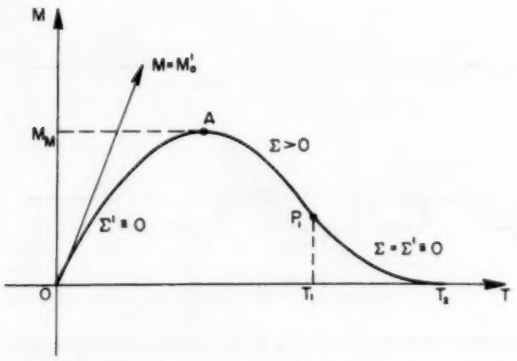


FIG. 19 OPTIMUM TRANSIENT FOR SYSTEM WITH DAMPING

The *maximum* value M_M of M is given by

$$M_M = \frac{M_0'}{\alpha} - \frac{\ln(1+\alpha M_0')}{\alpha^2} \quad \ldots\ldots [79]$$

From theory we can show that the transient is optimum in every sense and unique. The theory holds for transients that start at a point where $M_0 \neq 0$ as well as those where $M_0 = 0$.

SINGLE-LAG SYSTEMS

We now treat a physical system with a lag between servo and torque (in the prime-mover case). This lag is assumed in this section to be one associated with a pure time constant. The equation of the controlled system is now

$$\tau M'' + M' = C \quad \ldots\ldots [80]$$

instead of the first equation in Formulas [12]. Here τ is the time constant.

In the simplest prime-mover case the ideal transients are composed of one or two phases. This is no longer true for systems with lags. We introduce the function Σ given by

$$\Sigma = \psi + \frac{\{\psi'\}}{2} - (\operatorname{sgn}\psi')\tau^2 \ln^2\left\{1 + \sqrt{1-(1+[\operatorname{sgn}\psi']M'')e^{-|\psi'|/\tau}}\right\} \quad \ldots [81]$$

where

$$\psi = M + \tau M' \quad \ldots\ldots [82]$$

and

$$\begin{aligned} \operatorname{sgn}\psi' &= +1 \text{ if } \psi' > 0 \\ &= \ \ 0 \text{ if } \psi' = 0 \\ &= -1 \text{ if } \psi' < 0 \end{aligned} \quad \ldots\ldots [83]$$

The controlling function Σ for ideal transients involves two functions Σ_1 and Σ_2 where

$$\Sigma_2 = \psi + \frac{\{\psi'\}}{2} \quad \ldots\ldots [84]$$

and Σ_1 is the entire expression Σ given in Formula [81]. When Σ_1 becomes imaginary, we take the control function Σ equal to Σ_2. For an instant load rejection the two control functions are needed to bring one to equilibrium along the optimum curve. Here optimum is used in the same sense as before. The ideal transient for instant load rejection is shown in Fig. 20 for the one lag case.

Along the first phase OQ_1 of Fig. 20 we are controlling on the basis of Schedule [66] where

$$\Sigma = \Sigma_1 \quad \ldots\ldots [85]$$

Along this arc $\Sigma_1 > 0$ and $C' = -1$. At the point Q_1 the function Σ_1 becomes zero. At a point E after Q, we have $\psi' = 0$. Along the arc Q_1E

$$\Sigma_1 \equiv 0 \quad \ldots\ldots [86]$$

and $C' = +1$. Schedule [66] gives $C' = 0$ for this arc. However, after the point Q_1 the function Σ_1 will then become negative

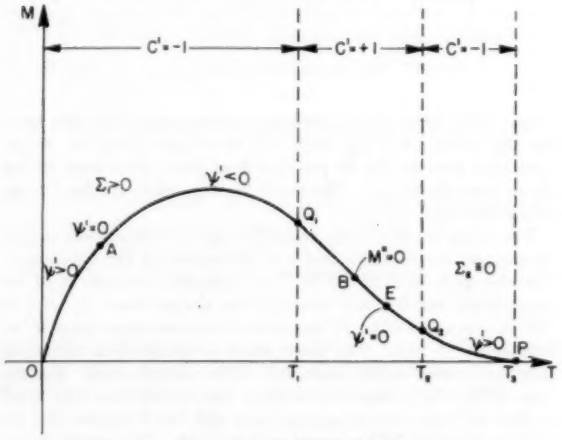


FIG. 20 IDEAL TRANSIENT FOR ONE-LAG CASE

immediately whence by Schedule [66] we have $C' = +1$. At the point E the function Σ_1 becomes imaginary. From the point E to Q_2 the function Σ_2 is negative and increasing. At the point Q_2 the function Σ_2 becomes zero. From E on we control on the basis of the function Σ_2. Schedule [66] with $\Sigma = \Sigma_2$ gives $C' = 0$ for the arc Q_2P. However, again with $C' = 0$ the function Σ will become positive immediately, whence by Schedule [66] we have $C' = -1$. We will now be brought to equilibrium (at least for practical purposes).

At the start we have $\psi' > 0$. This derivative decreases to zero at a point A. From the point A to the point E between Q_1 and Q_2 the quantity $\psi' < 0$. Between the points Q_1 and E there is the inflection point B, for which $M'' = 0$ occurs.

From the point E to the equilibrium point P the quantity ψ' is positive, increasing from the point E to the point Q_2 and decreasing from there to zero at the point P.

Equilibrium can be reached after any disturbance on the basis of Schedule [66] where control is based on the sign of Σ_1 or Σ_2 and during a transient the control function alternates (in the proper manner) between Σ_1 and Σ_2. Equilibrium can be reached on the basis of Schedule [66] in four or less phases. For "most" disturbances three phases suffice. *There are initial conditions for which the optimum transient is not one where the servo travels at full speed.* These would, however, be considered exceptional. Such cases arise when the M-curve is plunging toward the T-axis with a steep slope and reaching equilibrium with maximum servo speed would require the phases $C' = +1, -1, +1, -1$ in this order, or the phases $C' = -1, +1, -1, +1$ putting a hump in the M-curve that could otherwise be avoided.[4]

[4] The complete mathematical theory underlying the foregoing statements for the single-lag case of Formula [80] is quite complicated. It is in the files of the author, but has not been written up for publication.

Higher Lag Cases and Practical Compromises

It can be proved by mathematical considerations that for disturbances of a magnitude normally encountered the log term in Formula [81] is small compared to the rest of Σ, and can therefore be dropped. The controlling function is thus

$$\Sigma = \Sigma_2 \quad [87]$$

For a controlled system given by

$$\tau m'' + m' = K_1 c \quad [88]$$

and servo speed limitation

$$|c'| \leq K_2 \quad [3]$$

the formula for nonlinear control on the basis of Σ_2 becomes

$$c' = -K\left[(m + \tau m') + \frac{\{m' + \tau m''\}}{2K_1K_2}\right] \quad [89]$$

where K is as large as possible.

For a two-lag case we have the equation

$$\tau_1\tau_2 M''' + (\tau_1 + \tau_2)M'' + M' = C \quad [90]$$

with time constants τ_1 and τ_2. Equation [90] can be written as

$$M' = \frac{C}{(\tau_1 D + 1)(\tau_2 D + 1)} \quad [91]$$

where D stands for the derivative with respect to time T. The three controlling functions for this case are quite complicated and will be omitted, except for the analog of Σ_2 in Formula [84]. We let ψ be given by

$$\psi = M + (\tau_1 + \tau_2)M' + \tau_1\tau_2 M'' \quad [92]$$

For the last phase of a transient corresponding to a load rejection the quantity $\Sigma_2 \equiv 0$, where Σ_2 is as given in Equation [84], but with ψ as in Formula [92].

More generally, if the equation of the controlled system is given by

$$O(D)M' = C \quad [93]$$

for an operator

$$O(D) = (\tau_1 D + 1)(\tau_2 D + 1)\ldots(\tau_n D + 1) \quad [94]$$

in the derivative D, for each transient the function Σ_2 vanishes for the last phase, where Σ_2 is given as in Equation [84] and

$$\psi = O(D)M \quad [95]$$

Control on the basis of Σ_2 often gives a good approximation to the optimum transients. We can write Equation [93] as

$$\psi' = C \quad [96]$$

Control on the basis of Σ_2 and this ψ' *is control on the basis of an auxiliary variable;* namely ψ, instead of M. We make ψ come to equilibrium in an optimum manner. The ψ-system [96] is a no-lag system. When equilibrium is reached with $\psi \equiv 0$, the variable M comes to its equilibrium $M \equiv 0$ according to the differential equation

$$O(D)M = 0 \quad [97]$$

For the one-lag case of Equation [80] the variable M comes to equilibrium exponentially according to the law

$$M = M_s e^{-T/\tau}$$

where M_s is the value of M at the instant ψ reaches equilibrium.

If in Formula [92] the time constant τ_2 is small compared to τ_1 the Formula [92] can be replaced by the single-lag Formula [82] with $\tau = \tau_1$, or better, with $\tau = \tau_1 + \tau_2$. This is not the case if $\tau_1 = \tau_2$ (see Fig. 22).

If a dead time τ_d is introduced into Equation [93] so that this becomes

$$O(D)e^{-\tau_d D}M' = C \quad [98]$$

the dead time may be treated as a time constant unless it is large or dominates the other lags in the system (see Appendix 1 for some results on systems with dead time).

The foregoing theory applies if $O(D)$ has quadratic factors that do not factor further (in the field of real numbers), corresponding to second-order lags. Actually, $O(D)$ in Formula [95] may be any polynomial with positive coefficients.

If a damping term $f(M)$ is included on the left of Equation [93] this term can be dropped as in a previous section on systems with damping, since it improves the transients obtained on the basis of design without it. It is assumed, as in the introduction, that $f(M)$ is a monotonic nondecreasing function of M. Discussion of coulomb damping will be omitted here for the sake of brevity.

If now we have the equation

$$O_1(D)M = O_2(D)C \quad [99]$$

for the controlled system, where $O_1(D)$ and $O_2(D)$ are polynomials in D with the linear term missing in $O_2(D)$, all of the terms on the right of Equation [99] may often be dropped except the C-term and an equation of Type [93] used instead, at least when applying the nonlinear approach of this paper to obtain good, though not necessarily, optimum transients.

We remark that in the special case

$$M' = C' + C$$

it is impossible for M' and C to approach zero (equilibrium values) simultaneously while $C' = \pm 1$. It follows that the optimum transients in this case are not obtained by having the M-curve approach the T-axis with maximum $|C'|$.

Because of space considerations the treatment of the case where a term in the integral of M occurs on the left in Equation [99] will be omitted.

Where large lags are involved theory and experiments show that nonlinear control can be used to reduce servo jiggle that arises with linear control and give faster return to equilibrium.

Linear Band

With the control Formula [65] for Σ involving the absquare the system may be unstable. As M, M', etc., become numerically small in Σ their coefficients may be increased or decreased so that when $|M|$ and its derivatives are small, and one is thus near equilibrium, the control formula becomes

$$C' = -K\Sigma_L \quad [100]$$

for a sum such as

$$\Sigma_L = aM + bM' + eM'' \quad [101]$$

where a, b, and e are constants chosen so as to give good stability.

Practical Compromise

A promising approach for nonlinear control as discussed here is that where a system is treated as a one-lag system and the second derivative is dropped from Formula [89] in the interest of simplification. The control formula is now

$$c' = -K\left[m + \tau m' + \frac{\{m'\}}{2K_1K_2}\right] \quad [102]$$

where we use actual variables, rather than per unit quantities, the equation of the controlled system is given by Formula [88], and the servo limitation is given by the Relation [3]. To compensate for dropping the m''-term it is necessary to increase the coefficient of the absquare term so that Formula [102] is replaced by

$$c' = -K[m + (\tau + \beta|m'|)m'] \quad \ldots\ldots\ldots [103]$$

for a constant β. The coefficient of m' is now a variable, which tends to give a more highly damped response for large values of $|m'|$ than small ones. In practice the value of β is adjusted so that $\beta|m'|_{max}$ dominates τ for the maximum absolute value $|m'|$ of the derivative m' to be encountered in practice. Theory and experiments show that it is desirable to have $\beta|m'|_{max}$ equal to about 10τ or 25τ. The maximum value $|m'|_{max}$ depends on the magnitude of the disturbances encountered by the controlled system. An experimental transient for a case where Formula [103] is used is shown in Fig. 21. The top and bottom curves

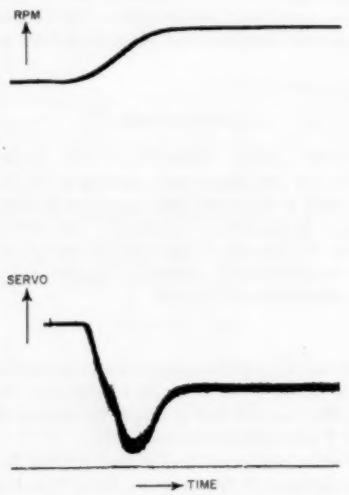


FIG. 21 EXPERIMENTAL NONLINEAR TRANSIENT FOR A CHANGE IN THE SETTING OF THE CONTROLLED VARIABLE

are for engine speed and servo position with respect to time. This figure shows the response to a change in the setting of the controlled variable.

If in Formula [103] we replace $|m'|$ by the "average" value it has for the larger disturbances we obtain a linear control that is at least a rough approximation to the nonlinear.

In practice it is convenient to choose K and τ so that the characteristic roots are optimum when the absquare term is dropped from Formula [103] and then to adjust β so that it is as large as possible without causing hunting of the controlled system after the worst disturbance is made to which the system is to be subjected. The use of absquaring as in Formula [103] allows one to have very poor characteristic roots when the derivative $|m'|$ is large, corresponding to fast movements of the servo, and ideal roots when near equilibrium, yielding optimum stability for small disturbances.

To allow for excessive disturbances it is desirable to bound the value of $|m'|$ or the term $|m'|m'$ in Formula [103].

DISCONTINUITIES

In the theory developed so far no allowance has been made for the fact that the servo stroke is limited. If the maximum servo speed is such that the servo goes through its stroke in a small fraction of the time required for the larger transients, so that application of the nonlinear theory gives essentially two-position control, the nonlinear theory developed here breaks down. Fortunately, this is not the case in many, if not most, applications. If the servo is not at the ends of its stroke during a major part of the transient, the nonlinear performance will normally not deviate too much from theory. If the servo speed is not the same for the two directions of travel of the servo, but is not too different for the two directions (theory and experiment indicate that a 2:1 variation is acceptable), the transients for nonlinear performance on the basis of this paper are still near optimum. The author's theory for the two-speed case will be omitted for the sake of brevity.

Tests of variations of all of the constants show that the nonlinear control treated here is not critical. Variations of 2:1 to 4:1 in the coefficients of Σ and the second power used in the absquare can be tolerated.

AREA OF USEFULNESS

If the lags in a system to be controlled are small, the range of disturbances is large, and the discontinuities are not severe, substantial improvement (by an order of magnitude) over existing control can be obtained for a range of disturbances with the nonlinear control described here. Experiments indicate that definite improvement can also often, if not generally, be obtained when the first two conditions are not satisfied.

A fundamental limitation, from the theoretical point of view, on the use of nonlinear control is that of noise. By "noise" we mean the unwanted part of the signal input to the controller. This signal contains the measurement of the controlled quantity, which is wanted, as well as another portion, which is not wanted. In the absquare $\{m'\}$ of m' the noise is worse than it is in m', and, in fact, may be almost as bad as in m''. On account of noise the gain constant K in Formula [65] is limited, if it is not limited for other reasons. It follows that for the nonlinear control to be effective the quantities m' (or ψ') must be large enough for the bigger disturbances to dominate the noise. *If a system is subject only to very small disturbances* such as those which cause M' to be 0.01 per cent maximum, *the noise may prevent the use of nonlinear control* involving the absquare.

If the noise is too great the servo jiggle will be excessive.

If a system has very large lags relative to the servo-stroke time, such as a ten-second time constant for a three-second servo, a term in M'' can be added to the ψ-term in Formula [84], and the coefficients in ψ can be changed to compensate for dropping the ψ'-term in Formula [84], so that the use of a linear Σ with second derivative is a satisfactory approximation to the performance that can be obtained by the nonlinear Σ of Formula [84], and the employment of a nonlinear Σ with the second derivative term in the absquare is not economically justifiable. However, Formula [103] still applies if the second derivative is not used.

For n sufficiently large, such as $n = 10$ or 100, a dead time τ_d may, for practical purposes, be replaced by n first-order lags in cascade with identical time constants equal to τ_d/n. The corresponding nonlinear theory requires the use of derivatives up to a high order, such as the 11th or more, which noise makes impossible. This difficulty is avoided in the precise dead-time theory of Appendix 1.

With a linear control function the initial overswing (or underswing) for a given instantaneous disturbance can be made the same as would be obtained for the optimum nonlinear control with the same servo, and bounded servo speed. However, nonlinear control with the absquare can be employed to improve the response to maximum disturbances by eliminating or reducing undesired oscillations (after the first swing), and to reduce over-

swings and underswings in the responses to lesser disturbances. For small disturbances a practical control is normally purely linear. Thus in the case of prime movers the overswing for an instantaneous 100 per cent load rejection may be adjusted to be the same for linear and nonlinear control, whereas for instant load rejections of less than 100 per cent the overswings are reduced. However, the linear control that gives the same overswing as for a satisfactory nonlinear control may be too oscillatory, so that even for 100 per cent load rejections the nonlinear control is to be preferred. *Essentially, by nonlinear control more efficient use is made of the servo.*

Co-operation of the prime-mover manufacturer in reducing lags in the system to be controlled can often result in substantial improvement in the control by permitting the use of a nonlinear governor nearer to that indicated by the theory for the no-lag case. Thus in one application the time constant of a servo (supplied by the governor user) due to trapped air was $^1/_3$ sec, when it should have been about $^1/_{100}$ sec. The result was a different order of magnitude in the speed deviation for the maximum disturbance to be encountered.

Response to Sinusoidal and Other Disturbances

Consider the equation

$$M' = C - \sin \omega T$$

Optimum performance is obtained by letting

$$C = \sin \omega T$$

This condition can be achieved by linear control only for discrete values of ω, and in no case by the nonlinear control of this paper. Frequency-response runs on systems based on optimum nonlinear control and on the best linear control show about the same results except that at large amplitudes of the forcing sine wave the nonlinear control yields a lower resonant frequency where the system response has a peak.

For linear throttle bursts (constant velocity) of 250 to 1000 rpm in $^1/_2$ to 3 sec on a simulated 3000-rpm airplane engine, connected to a simulated propeller the overspeeds for nonlinear control based on Formula [103] were about half of the overspeeds obtainable by linear control. In this example the servo lag was taken to be 0.02 sec (time constant).

Space does not permit the inclusion of the author's theory for throttle bursts and other kinds of disturbances.

Physical Nonlinear Equipment

Methods of producing the mathematical operations needed to yield the nonlinear control functions of this paper are well known. Thus the use of d-c circuits for differentiating electrically and dashpots for doing this mechanically is classical.

The use of absquaring physical components is also classical. Nonlinear resistors exist for which the current is proportional to the absquare of the voltage. The pressure drop across a sharp-edged orifice is proportional to the absquare of the flow. Thus the absquare is easy to produce by physical devices in common use. However, the improvement of the transients in a problem must be weighed against the cost of including nonlinear terms in the control function.

Experimental Results

All of the points of the theory in this paper have been checked experimentally in the laboratory of the author's company by electronic and other means on simulated and actual engines.

Some experimental results are shown in Figs. 22 and 23. Fig. 22(a) is for a one-lag ($^1/_5$-sec time constant) system where Formula [89] is used for control. This figure shows the response to instant load rejections and increases. The curves in Fig. 22(b) are the same as for Fig. 22(a) except that the lag with $^1/_5$-sec time constant has been split into two identical lags with $^1/_{10}$-sec time constant. Fig. 22(c) is the same thing with the $^1/_5$-sec lag split into three $^1/_{15}$-sec lags. The asymmetry of the rpm swings in Fig. 22(c) is due to minor errors in the equipment and is to be disregarded. The vertical scales for Figs. 22(b) and 22(c) differ a little from those for Fig. 22(a). For Figs. 22(b) and 22(c) the time constants are lumped into one in the control Formula [89]. To eliminate the oscillations near equilibrium in Fig. 22(c) one must employ a sizable linear band about equilibrium or a control formula with higher derivatives, or some other technique such as a deadband. In Fig. 23 is shown the rpm response of an engine to a load increase and a load rejection for the case where the engine and filter in the system have four lags, with time constants 0.4 sec, 0.1 sec, 0.1 sec, and 0.05 sec, respectively, and a control formula of Type [89] with second derivative is used. Here the 0.4-sec lag dominates the others.

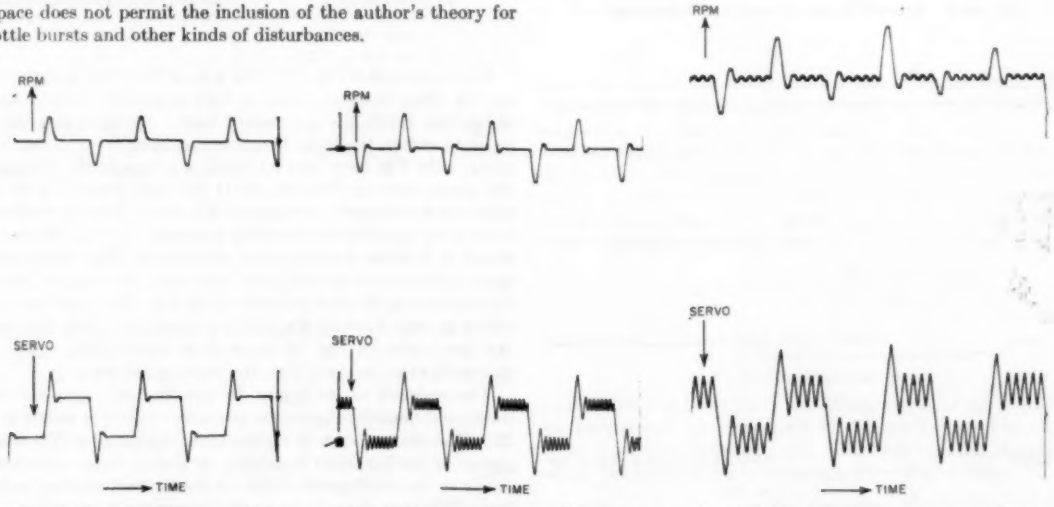


Fig. 22(a) Experimental Response for a Single-Lag System With a Nonlinear Control

Fig. 22(b) Response of Engine With Two Equal Lags

Fig. 22(c) Response of Engine With Three Equal Lags

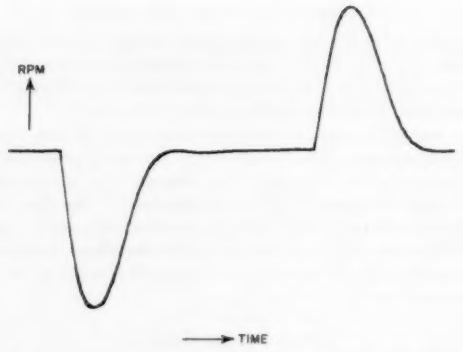


FIG. 23 RESPONSE OF SYSTEM WHERE ONE LAG DOMINATES THREE OTHERS

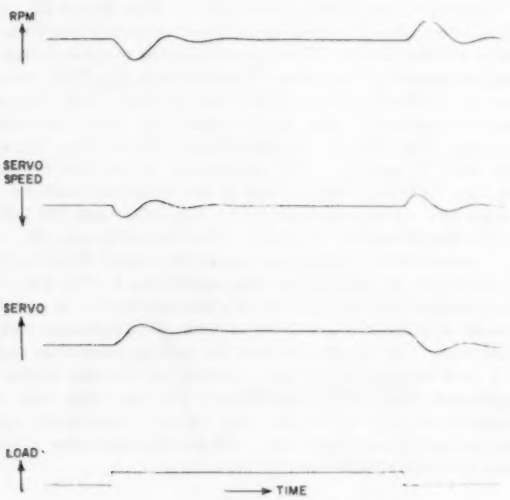


FIG. 24(*a*) LINEAR RESPONSE TO LOAD REJECTION

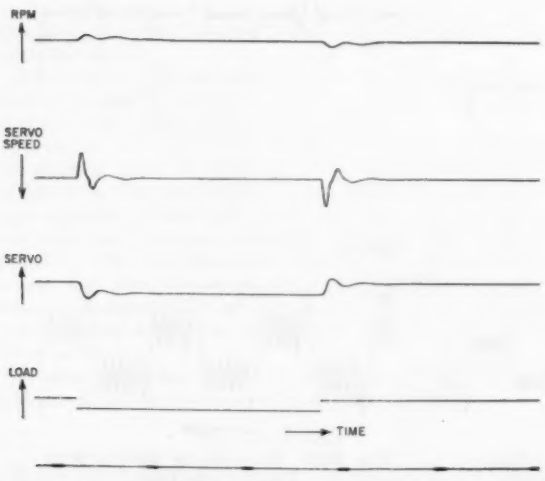


FIG. 24(*b*) NONLINEAR RESPONSE TO SAME DISTURBANCE

In Figs. 24–27 are shown the experimental results of using an absquare term (Formula [103]) for prime-mover systems where the disturbances are so small that *the servo speed never reaches a maximum value*. Fig. 24(*a*) gives the response of such a system to load disturbances where the governor is a linear one adjusted to give the best results that can be obtained with a linear control function where servo speed c' is a linear combination of speed deviation m and acceleration m'. In this figure load is taken on and then rejected. In Fig. 24(*b*) is shown the response of the same system to the same disturbances when a practical amount of absquaring of the acceleration m' is introduced; i.e., Formula [103] is employed. Note the improvement in speed deviation by an order of magnitude. Note also that the part of the transient after the first swing is affected very little because the absquare plays a small role when one is near equilibrium. In Fig. 24(*b*) the case of load rejection is shown first, whereas in Fig. 24(*a*) it is shown last.

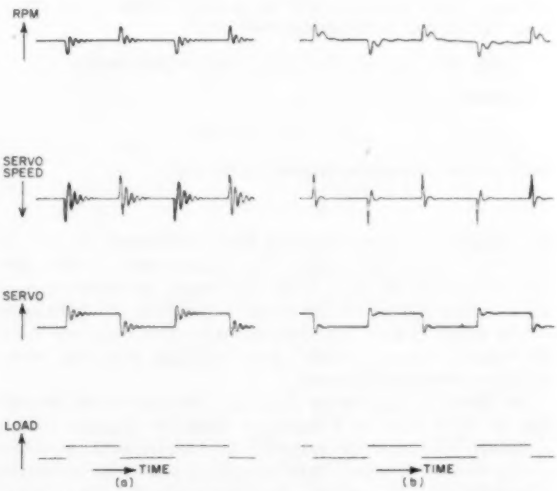


FIG. 25(*a*) LINEAR RESPONSE TO DISTURBANCE

FIG. 25(*b*) NONLINEAR RESPONSE TO SAME DISTURBANCE WITH SAME SPEED DEVIATION

For the curves of Fig. 25(*a*) the gain of the linear governor used for Fig. 24(*a*) has been raised as high as possible without having the system break into a sustained hunt. By increasing the gain so as to obtain the same linear performance (i.e., characteristic roots) as in Fig. 24(*a*) and introducing a term in the absquare of the acceleration in Formula [103] the same overswing for load rejection is obtained, as shown in Fig. 25(*b*), but the oscillations have been removed for practical purposes. In Fig. 26 the constants in a linear governor [the governor of Fig. 24(*a*)] for the same prime-mover system have been adjusted to give the same rpm overswing for load rejection as in Fig. 25(*a*) and rpm transients as near those of Fig. 25(*b*) as possible. Note how slowly the rpm curve of Fig. 26 drags in to equilibrium. Note also the oscillations, evident from the servo speed trace.

The response of an engine to quarter-load rejections under linear and nonlinear hydraulic governor control is shown in Fig. 27. The linear governor represented by Equation [70] was adjusted to the border of instability so that a slight numerical increase in the coefficients of the control formula resulted in hunting. This was done to make the overswing a minimum. The nonlinear governor, based on Formula [103], was adjusted so that a good transient was obtained for Fig. 27 without encounter-

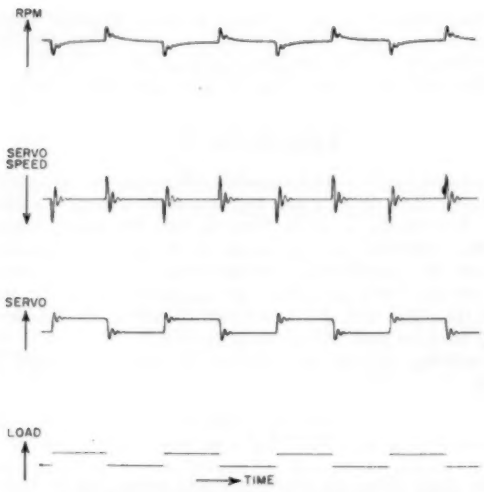


FIG. 26 LINEAR RESPONSE INITIALLY SIMILAR TO NONLINEAR RESPONSE

ing instability. Note that the area under the spike for the nonlinear case is about half of what it is for the linear. For this test the linear and nonlinear governors were actually the same governor except for the physical components generating the absquare in the nonlinear case and the derivative term m' in the linear.

Thus even though *the servo speed does not attain its maximum value* for the disturbances under consideration, the introduction *of the absquare can be used to substantially improve the response* to the "larger" of the disturbances to be encountered.

SUMMARY

In unpublished work the author has treated such topics as the effect of deadband, changing gain in reaching a linear control band about equilibrium, the use of linear and nonlinear approximations to nonlinear control functions, the effect of a second sudden disturbance before the first has died out, the employment of different types of linear control zones near equilibrium, the use of an arbitrary $g(c)$, the determination of control functions for higher-order systems, and other topics which need to be covered in a complete treatment of the subject, but which will be omitted here.

Instant load rejections and the corresponding transients are of considerable concern to the user of speed governors. Let $O(D)$ be a polynomial in D with positive coefficients. To obtain optimum transients in the case

$$O(D)m' = K_1c - l \quad [104]$$

where

$$|c'| \leq K_2 \quad [3]$$

it is necessary to use control functions which include

$$\Sigma = \psi + \frac{\{\psi'\}}{2K_1K_2} \quad [105]$$

where $\{\psi'\}$ is the absquare $|\psi'|\psi'$ of ψ' and

$$\psi = O(D)m \quad [106]$$

Since Σ in Formula [105] involves the third or higher derivatives of m if $O(D)$ is of the second order or higher, due to noise

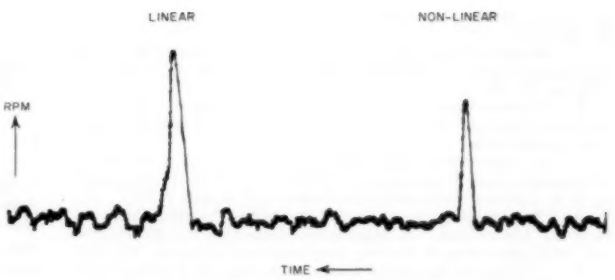


FIG. 27 PARTIAL-LOAD REJECTIONS ON ENGINE

precise optimum transients cannot be realized physically for higher-lag systems.

Good transients can often be obtained by using

$$c' = -K\Sigma \quad [107]$$

as the only nonlinear control formula, where K is as large as possible, and a linear band, dead zone, or something else is employed to achieve stability where necessary. If the equation is

$$m' + f(m) = \frac{1}{O(D)} K_1c - l \quad [108]$$

with a damping term $f(m)$ that does not decrease as m increases, this term may be dropped and good results obtained by applying the theory for Equation [104]. Let ψ denote $m + \tau m'$. In the case where $O(D)$ is the expression $(\tau D + 1)$ only one control function

$$\Sigma = \psi + \frac{\{\psi'\}}{2K_1K_2} - (\operatorname{sgn}\psi')\tau^2K_1K_2 \ln^2\left\{1 + \sqrt{1 - \left(1 + [\operatorname{sgn}\psi']\frac{m''}{K_1K_2}\right)e^{\frac{-|\psi'|}{\tau K_1K_2}}}\right\} \quad [109]$$

is needed to give optimum transients where $\{\psi'\}$ is the absquare of ψ'. When the log term in Formula [109] is imaginary it is dropped.

Formula [105] may often be replaced by

$$\Sigma = m + (\tau + \beta|m'|m') \quad [110]$$

especially where the servo will not attain its maximum speed. Formula [110] may frequently in turn be replaced by a linear expression by using an average value of $|m'|$.

When higher derivative terms in c are introduced on the right of Equation [104] the function Σ of Formula [105] can still be used. If a term in c' is introduced on the right in Equation [104] the nonlinear theory of this paper breaks down. If dead time τ_d is introduced into Equation [104]

$$\Sigma = \psi \pm \tau_d^2K_1K_2 + \frac{\{\psi' \mp 2\tau_dK_1K_2\}}{2K_1K_2} \quad [111]$$

may be used to treat large disturbances.

ACKNOWLEDGMENT

The author is indebted to Mr. William I. Caldwell of the Taylor Instrument Companies, Mr. G. Forrest Drake of the Woodward Governor Company, and others for valuable suggestions which were used to improve the exposition of this paper.

BIBLIOGRAPHY

1 "Nonlinear Techniques for Improving Servo Performance,

by Donald McDonald, *National Electronics Conference*, vol. 6, 1950, pp. 400–421.

2 "Multiple Mode Operation of Servomechanisms," by Donald McDonald, *Review of Scientific Instruments*, vol. 23, no. 1, January, 1952, pp. 22–30.

3 "A Phase-Plane Approach to the Compensation of Saturating Servomechanisms," by A. M. Hopkin, Trans. AIEE, vol. 70, part I, 1951, pp. 631–639.

4 "A Topological and Analog Computer Study of Certain Servomechanisms Employing Nonlinear Electronic Components," by R. C. Lathrop, PhD thesis, University of Wisconsin, 1951.

5 "The Stabilization of On-Off Controlled Servomechanisms," by A. M. Uttley and P. H. Hammond, *Automatic and Manual Control*, 1952 publication of the 1951 Cranfield Conference, pp. 285–307.

6 "Differential Equations With a Discontinuous Forcing Term," by D. W. Bushaw, Report No. 469, January, 1953, Experimental Towing Tank, Stevens Institute of Technology, Hoboken, N. J., PhD thesis, Dept. of Math., Princeton University, Princeton, N. J.

7 "Predictor Servomechanisms," by L. M. Silva, Trans. IRE, vol. CT-1, no. 1, March, 1954, pp. 56–70.

8 "Nonlinear Optimization of Relay Servomechanisms," by L. M. Silva, University of California Institute of Engineering Research, series no. 60, issue no. 106, April 15, 1954.

9 "Predictor Control Optimalizes Control-System Performance," by L. M. Silva, Trans. ASME, vol. 77, 1955, pp. 1317–1323.

10 "An Investigation of the Switching Criteria for Higher Order Servomechanisms," by Irving Bogner and L. F. Kazda, Trans. AIEE, vol. 73, part II, Applications and Industry, 1954, pp. 118–127, Paper 54–44.

11 "Differential Equations," by R. P. Agnew, McGraw-Hill Book Company, Inc., New York, N. Y., 1942.

12 "Optimum Switching Criteria for Higher Order Contactor Servo With Interrupted Circuits," by S. S. L. Chang, AIEE 55–549.

13 "Analysis and Design Principles of Second and Higher-Order Saturating Servomechanisms," by R. E. Kalman, AIEE 55–551.

14 "Errors in Relay Servomechanisms," by L. F. Kazda, Trans. AIEE, vol. 72, part II, Applications and Industry, 1953, pp. 323–328.

15 "Effects of Friction in an Optimum Relay Servomechanism," by T. M. Stout, Trans. AIEE, vol. 72, part II, Applications and Industry, 1953, pp. 329–336.

Appendix 1

We shall consider a system with dead time given by the equation

$$M' = e^{-\tau_d D}C - L \quad [112]$$

with

$$|C'| \leq 1 \quad [113]$$

Suppose that

$$L \geq (\sqrt{2} - 1)\tau_d \quad [114]$$

If the load L is suddenly dropped at $T = 0$ an optimum transient is obtained where one goes through three phases corresponding to $C' = -1, +1, 0$. Before the disturbance

$$C = L, C' = 0$$

The change from $C' = -1$ to $C' = +1$ must occur when the control function Σ given by

$$\Sigma = M + \tau_d{}^2 - \frac{(M' - 2\tau_d)^2}{2} \quad [115]$$

is zero. From $T = 0$ to $T = \tau_d$ we have $M' = L$ and $M = LT$. If L is sufficiently large, as when $L \geq 2\tau_d$, we have $\Sigma > 0$ and $\Sigma' < 0$ before the instant when C' changes from -1 to $+1$. When

$$M = \frac{\tau_d{}^2}{2}, M' = -\tau_d \quad [116]$$

we must switch to $C \equiv 0$. The C-curve reaches equilibrium τ_d units before this happens for M.

Corresponding results are obtained if load is suddenly taken on. For large enough L relative to τ_d we can replace the last term in Formula [115] by the absquare term $\{M' - 2\tau_d\}/2$.

Modifications of the theory apply if L is small relative to τ_d.

Appendix 2

We consider a system with Equation [93] subject to the Condition [113]. Here $O(D)$ is any polynomial with positive coefficients. Introducing ψ as in Formula [95] we derive Equation [96]. Consider the last phase of an optimum transient that leads M to equilibrium. Suppose that $C' = +1$ for this phase, whence $C \equiv 0$ thereafter. We cannot have $\psi = 0$ at the start of the phase since $\psi'' = +1$ until the end of the transient when ψ becomes zero. Thus the shift from $C' = -1$ to $C' = +1$ in entering the last phase (before $M = 0$) occurs when Σ given by

$$\Sigma = \psi + \frac{\{\psi'\}}{2} \quad [117]$$

becomes zero. Thus *the absquare occurs in the last switching function.*

The foregoing treatment holds equally well when $C' = -1$ before equilibrium.

If Equation [93] is replaced by

$$O(D)M' = e^{-\tau_d D}C \quad [118]$$

we let

$$U = e^{-\tau_d D}C$$

whence Equation [118] goes into

$$O(D)M' = U \quad [119]$$

where Relation [113] implies that

$$|U'| \leq 1 \quad [120]$$

We can write ψ for $O(D)M$ and obtain

$$\psi' = U \quad [121]$$

The argument at the beginning of this Appendix applies except that C' must switch values τ_d units of time before it would if the dead time were absent (see Appendix 1). For sufficiently large load rejections or acceptances the final phase before equilibrium (i.e., the last phase for which $\psi'' \neq 0$) is initiated when the control function

$$\Sigma = \psi \pm \tau_d{}^2 + \frac{\{\psi' \mp 2\tau_d\}}{2} \quad [122]$$

becomes zero. The absquare thus occurs in the treatment of a system with the Equation [118]. The top signs in Formula [122] apply to load rejections and the bottom to load acceptances.

Discussion

M. J. Nowak.[5] The author develops the absquare control function, with particular application to the optimum governor speed control of an engine. In general the equation for this type of servo application is

$$O(D)M' = C \quad [123a]$$

$$|C'| \leqslant 1 \quad [123b]$$

This general equation represents a large class of servo applica-

[5] Fellow, Engineering Mechanics Division, Stanford University, Stanford, Calif.

tions, for which a particular interpretation is the case where M is an engine-speed deviation and C is the position of a controlling governor; the characteristic of the governor is that its maximum speed is limited (e.g., this may represent the rate of opening of a fuel valve).

For this control situation the intuitive idea of having the servo travel at maximum speed to obtain optimum transient response is used and justified by the author to develop the absquare control function Σ for the case where engine time lags are negligible so that $O(D) = 1$

$$\Sigma = M + {}^1/_2 \, |M'|M' \quad \ldots\ldots\ldots\ldots [124]$$

In general a control function is a function of the process error and its time derivatives, whose sign determines the direction in which the servo is moving (at maximum speed). These required properties of the motion are sensed and combined into the control function; then the times at which the control function is zero determine the switching points for the servo.

It turns out that absquare control has use not only in the simple case analyzed by the author but also in the general Equation [123]. This more general case will be developed briefly for arbitrary engine time delay and arbitrary initial conditions of engine speed. The result is that even in this general situation the control function has the form

$$\Sigma = M + {}^1/_2 \left| \frac{M'}{C'} \right| M' + 2TM' - 2C'T\hat{T} \ldots [125]$$

Thus the exact switching function contains corrections to the absquare function of first and second order in the equivalent time constant of the engine.

For such step switchings involved in this type of discontinuous control the Heaviside transformation theorem is useful; for a differential equation which relates an output y to a step input x

$$y + a_1 y^{(1)} + a_2 y^{(2)} + \ldots + a_n y^{(n)} = x \ldots\ldots [126a]$$

$$O(D)y = x \qquad O(D) = (1 + T_1 D) \ldots (1 + T_n D) \quad [126b]$$

the usual Heaviside transformation is

$$y(t) = y_s + x \sum_{k=1}^{n} \frac{-T_k}{O'\left(\frac{-1}{T_k}\right)} e^{-t/T_k}$$

$$y_s = [y(t)]_{t=\infty} = \left[\frac{x}{O(D)}\right]_{D=0} \ldots\ldots [127]$$

The limitation in using this formula is that it applies only to the situation where there is no motion prior to applying the step input, whereas in discontinuous switching control the servo may be reversed several times under arbitrary conditions. However, these arbitrary initial conditions can be taken into account by using operational methods. Thus the original equation can be integrated once by using the operator $\bar{D} = 1/D$ and applying the initial conditions

$$\bar{D}x = \bar{D}y + a_1 y + a_2 y^{(1)} + \ldots + a_n y^{(n-1)} - (a_1 y_0 + a_2 y_0^{(1)} + \ldots + a_n y_0^{(n-1)}) \ldots\ldots [128a]$$

This process can be repeated for n-integrations to obtain

$$\left.\begin{aligned} \bar{D}^n x = {} & \bar{D}^n y + a_1 \bar{D}^{n-1} y + a_2 \bar{D}^{n-2} y + \ldots + a_n y \\ & - (a_n + a_{n-1}\bar{D} + \ldots + a_1 \bar{D}^{n-1}) y_0 \\ & - (a_n \bar{D} + a_{n-1}\bar{D}^2 + \ldots + a_2 \bar{D}^{n-1}) y_0^{(1)} \\ & - \ldots \\ & - a_n \bar{D}^{n-1} y_0^{(n-1)} \end{aligned}\right\} \ldots [128b]$$

By now applying the operator D^n the result can be put in the following form

$$Oy = x + O_0 y_0 + O_1 y_0^{(1)} + \ldots + O_{n-1} y_0^{(n-1)} \ldots [129]$$

$$\left.\begin{aligned} O(D) &= 1 + a_1 D + a_2 D^2 + \ldots + a_n D^n \\ O_0(D) &= (a_1 + a_2 D + \ldots + a_n D^{n-1}) D \\ O_1(D) &= (a_2 + \ldots + a_n D^{n-2}) D \\ O_{n-1}(D) &= a_n D \\ O_k(D) &= \sum_{i=k+1}^{n} a_i D^{i-k} = \frac{1}{D^k}\left[O(D) - \sum_{i=0}^{k} a_i D^i\right] \end{aligned}\right\} \ldots [130]$$

The output is now produced not only by the applied step but also by steps due to the initial conditions. The usual Transformation [127] can be applied to each step to obtain the complete output; thus the general Heaviside transformation for arbitrary initial conditions is

$$y = y_s - \sum_{k=1}^{n} \frac{T_k e^{-t/T_k}}{O'\left(\frac{-1}{T_k}\right)} \left[x + \sum_{i=0}^{n-1} O_i\left(\frac{-1}{T_k}\right) y_0^{(i)}\right] \ldots\ldots [131]$$

To facilitate application of the formula the following relations are useful

$$\left.\begin{aligned} O\left(\frac{-1}{T_1}\right) &= 0 \\ O'\left(\frac{-1}{T_1}\right) &= \frac{1}{T_1^{n-2}} (T_1 - T_2) \\ &\qquad (T_1 - T_3) \ldots (T_1 - T_n) \\ O_0\left(\frac{-1}{T_1}\right) &= O\left(\frac{-1}{T_1}\right) - 1 = -1 \\ O_1\left(\frac{-1}{T_1}\right) &= -(T_2 + T_3 + \ldots + T_n) \\ O_{k+1}\left(\frac{-1}{T_1}\right) &= -T_1 O_k\left(\frac{-1}{T_1}\right) - a_{k+1} \end{aligned}\right\} \ldots [132]$$

For the servo application [123] the formula can be applied to the equation

$$O(D)M'' = C'$$

where C' is a step equal to ± 1

$$\left.\begin{aligned} M'' = {} & C' - \sum_{k=1}^{n} \frac{T_k e^{-t/T_k}}{O'\left(\frac{-1}{T_k}\right)} \left[C' + \sum_{i=0}^{n-1} O_i\left(\frac{-1}{T_k}\right) M_0''^{(i)}\right] \\ & M'' = C' - \tilde{M}'' \end{aligned}\right\} \ldots [133]$$

The transient term on the right may be denoted by $\tilde{M}''$ and the equation integrated twice with initial conditions applied

$$(M + \tilde{M}) = (M_0 + \tilde{M}_0) + (M_0' + \tilde{M}_0')t + \frac{1}{2}(M_0'' + \tilde{M}_0'')t^2 \ldots\ldots [134a]$$

$$(M' + \tilde{M}') = (M_0' + \tilde{M}_0') + (M_0'' + \tilde{M}_0'')t \ldots [134b]$$

This form of the equation illustrates the author's comment that

absquare control for the general case can be applied to the fictitious speed $\hat{M} = M + \tilde{M}$ instead of the actual engine speed. Moreover, this formula can be used to develop a switching criterion for the actual engine speed in this general case.

The general method to determine the switching criterion which leads the system to equilibrium with zero error is to find the time t_1 (in terms of the initial conditions M_0 and M_0') such that the servo forcing can be removed (stepped to zero) and the engine will coast to a final speed with no speed error. For a system with no engine time lags this condition on t_1 is that the speed error and acceleration are zero at t_1; for an engine with time delays the condition is that $M + \tilde{M}$ and $M' + \tilde{M}'$ must be zero at t_1. This important condition can be obtained readily from the general Heaviside transformation applied to the final switching to $C' = 0$.

It turns out that the simplified cases in which the engine time lags are neglected, or a single time lag is included, are insufficient to reveal the essential form of the switching criterion.

In the general case an exact solution can be carried through in terms of two equivalent time constants T, $\hat{T}$ of the order of the engine time constant, which are obtained from the general Heaviside transformation

$$\tilde{M}_0' = -2C'T \qquad \tilde{M}_0 = 2C'T(T - \hat{T}) \ldots\ldots [135]$$

To indicate the development of these equivalent time constants consider first the cases of one and two engine time delays with switching points far enough apart that the engine can settle into its final state of forced motion at $M^r = C'$. In these cases the following substitutions apply

$$\left[C' + \sum_{i=0}^{n-1} O_i \left(\frac{-1}{T_i}\right) M_0^{r(i)}\right] = 2C' \ldots\ldots [136]$$

One delay

$$\tilde{M}' = -2C'T_1e^{-t/T_1} \qquad \tilde{M}_0' = -2C'T_1$$

$$\tilde{M} = 2C'T_1^2e^{-t/T_1} \qquad \tilde{M}_0 = 2C'T_1^2$$

$$T = T_1 \qquad \hat{T} = 0 \qquad T\hat{T} = 0$$

Two delays

$$\tilde{M}' = \frac{-2C'}{T_1 - T_2}\left[T_1^2e^{-t/T_1} - T_2^2e^{-t/T_2}\right] \qquad \tilde{M}_0' = -2C'(T_1 + T_2)$$

$$\tilde{M} = \frac{2C'}{T_1 - T_2}\left[T_1^2e^{-t/T_1} - T_2^2e^{-t/T_2}\right] \qquad \tilde{M}_0 = 2C'(T_1^2 + T_1T_2 + T_2^2)$$

$$T = T_1 + T_2 \qquad \frac{1}{\hat{T}} = \frac{1}{T_1} + \frac{1}{T_2} \qquad T\hat{T} = T_1T_2$$

In this method of equivalent time constants C' is the maximum servo speed (± 1) if the switching points are far enough apart; if the switching points are close together C' may have some modified value obtained from Equation [136], depending on the previous switching. In any case Equation [134b] can be solved for t_1

$$t_1 = -\frac{1}{C'}(M_0' - 2C'T) \ldots\ldots [137]$$

This value of t_1 can be substituted into Equation [134a] and the exact equation for the switching criterion is obtained

$$\Sigma = M + \tfrac{1}{2}\left|\frac{M'}{C'}\right| M' + 2TM' - 2C'T\hat{T} \ldots [125]$$

The equivalent parameters T, $\hat{T}$, C', which can be estimated from the general Heaviside transformation, actually depend on the time delays of the engine, the time between switching points, and the initial conditions of the previous switching; generally it may be too expensive and also unimportant to sense this complicated dependence exactly, and experimental coefficients can be used in the switching criterion. Furthermore, since the constant correction to the switching function is a second-order effect this general analysis supports the author's practical switching criterion

$$\Sigma = M + (T + \beta|M'|)M' \ldots\ldots [138]$$

It appears that if a more accurate switching criterion is warranted for large engine time delays a constant bias also should be included in the control function.

T. M. Stout.[6] Many of the previous papers on optimum nonlinear control systems, including those by the writer, have described theoretical or analog computer studies or, in some cases, tests performed on experimental instrument servomechanisms. The author must be among the first to try these concepts in real systems, and his paper is therefore a genuine contribution.

Familiarity may breed a mistaken notion of the complexities of a subject. The writer, nevertheless, considers the statements "The mathematics of this theory is far beyond the limits of this paper" and "this theory is extremely complicated" to be unwarranted. The only permissible values of the manipulated variable that need to be considered are $+1$ and -1. In second-order systems, the corresponding response curves can be plotted in a phase plane. It is easy to visualize what combination of these curves is needed to reach equilibrium from any set of initial conditions and to determine the necessary control or switching function. The resulting transients can be shown to be optimum by making slight alterations in the switching procedure and showing that these increase the response time. For third-order systems, corresponding arguments can be carried out in a three-dimensional phase-space. This essentially graphical but perfectly rigorous approach is extended with difficulty to fourth or high-order systems, because of our inability to draw the multidimensional figures required, but the line of reasoning is still applicable. The strictly analytical approach appears unduly cumbersome, even for the simplest systems, so that the phase-plane or phase-space attack seems to be advantageous in all cases.

The paper contains a number of statements and asides which might serve as material for several subsequent papers. The author's remark, "There are initial conditions for which the optimum transient (in the single-lag case) is not one where the servo travels at full speed," deserves elaboration, as does the special case

$$M' = C' + C$$

The discussion of discontinuities, two-speed systems, dead time, dead band, approximations to the ideal control functions, and the effects of multiple, sinusoidal, and other disturbances likewise could be expanded. A question of some importance, not mentioned here, is the effect of intermittent operation on the power rating and life of the actuating device or servo. These matters require investigation before practical applications of optimum nonlinear control systems can be made.

"Compromises to preserve performance but reduce cost" appear essential, and the author correctly stresses the joint importance of performance, engineering, and economic considera-

[6] The Ramo-Wooldridge Corporation, Los Angeles, Calif.

tions. One cannot help wondering whether his optimism concerning the potentialities of optimum nonlinear control is influenced by an apparent, but possibly misleading, simplicity of the governor problems. In the cases discussed, low-order differential equations seem to be adequate to describe the systems, and disturbances of interest seem to be well approximated by step functions. Application of these control concepts to more complicated systems, described by higher-order equations and subjected to less specialized inputs and disturbances, seems less promising. The author would doubtless agree that the theory of optimum nonlinear control should serve only as a guide in the employment of deliberate nonlinearities and should not be applied blindly for its own sake.

T. J. Higgins.[7] This paper deals with what are referred to—at least in electrical engineering—as relay-type (or off-on) servomechanisms. The author cites a number of the principal papers that have been published on the analysis and design of such systems. A substantially equal number of yet other papers on these systems is to be found in the discusser's exhaustive bibliography on nonlinear control systems.[8]

To these many writings on relay-type systems the paper under discussion comprises a most valuable addition. First, the discusser—who has read all of the mentioned published literature on relay servomechanisms—found it to be very clearly written, which is not the case for all the earlier published papers. The theory is developed in considerable detail, so that all points of the theoretical developments are easily understood and grasped. Also, the author has adopted the very excellent procedure of concentrating attention on several of the simpler cases and has examined all possible initial modes of operation of each. Such procedure enables the reader to gain a clear insight into the physical actions through pertinent interpretation of the corresponding mathematical analysis. It also renders very clear (to the qualified reader!) the general procedure to be followed in analyzing more complicated systems such as some of those mentioned, but not taken up by the author, and reveals some of the essential physical phenomena that must be dealt with in designing nonlinear control systems, without obscurement by heavy mathematical manipulation.

Finally, and on the practicing side, the advance of experimental data in confirmation of the analytical work evidences the correctness of the author's work and provides substantial evidence of the great values of analytic procedures for determining optimum performance of complicated nonlinear systems.

In conclusion, the writer would extend to the author his sincere appreciation of the pleasure afforded him by the reading of this well-wrought and lucidly written paper on a difficult and currently interesting phase of control theory.

Author's Closure

The author is in general agreement with the points raised by Mr. Nowak. The theme of Mr. Nowak's discussion is that the absquare used by the author to treat systems without lags also applies to systems with lags. This was actually brought out in the paper in connection with Equation [93]. In fact, Mr. Nowak's equation

$$M'' + \tilde{M}'' = C' \quad \text{[139]}$$

obtained from Relations [133] is identical with the equation

$$O(D)M'' = C'$$

obtained from Equation [93] by differentiation, and the switching function involving the absquare has already been derived in the paper. This function is given by Σ in Formula [26] when M is replaced by ψ where

$$\psi = M + \tilde{M}$$

and is identical with the function Σ in Formula [125]. The author notes that $(M_0'' + \tilde{M}_0'')$ in Equation [134a] is C', and hence is equal to ± 1. His treatment of Equation [96], like that of Mr. Nowak, involves control on the basis of a fictitious speed ψ, i.e., $M + \tilde{M}$ (denoted by $\hat{M}$ in the discussion), and after ψ attains the equilibrium condition $\psi \equiv 0$, the engine will coast to the equilibrium value $M = 0$. The theory given in the discussion is thus the same as that of the author except for the point of view.

The author disagrees with the derivation of Equation [129], which is not mathematically rigorous. Starting with Equation [126a] by integration and differentiation Mr. Nowak derives Equation [129], which is the same as Equation [126a] except that terms have suddenly appeared on the right. From a given equation one cannot derive a new equation that contains more than at the start. Heaviside's work itself was lacking in rigor. This invalidated some of his findings, but did not detract from the over-all value of his work. The same thing can be said about the discussion under consideration.

Dr. Stout possesses a keen and thorough knowledge of the field in which the paper is written, and his remarks are well taken. He states that only the values ± 1 of the servo speed C' need be considered in the argument. This is true if one restricts the theory to on-off servomechanisms. Workers in the field generally have taken this position. If one restricts oneself to the on-off case it is a simple matter to establish whether or not a given transient is optimum in the sense of the paper. In the theory developed by the author it is assumed only that C' is between -1 and $+1$. Since a class of transients, optimum in the sense of this paper, requires that the servo speed C' be in absolute value between 0 and 1, at least during part of the transient, it is not sufficient to consider operation at saturation only.

Although in industry one is primarily concerned with the production of economical engineering designs, regardless of how these designs are obtained, it is nevertheless valuable to have the analytical treatment at hand. The analytical, that is rigorous, mathematical treatment of a theory is absolutely necessary for an understanding of the applications of this theory. This is a point of view that the author knows Dr. Stout shares with him.

The treatment of the case

$$M' = \tau C' + C \quad \text{[140]}$$

for a constant τ is not complicated and will be given here. It is assumed that $|C'|$ is bounded by 1, as in Relations [12]. Suppose that at $T = 0$ we have the initial conditions

$$M = M_0, \quad M' = M_0', \quad C' = 0, \quad C = M_0'$$

The best transient is obtained by letting

$$C' = -(\operatorname{sgn} M)$$

when $M \neq 0$, and

$$C' = -\frac{C}{\tau}$$

when $M = 0$. Note that if $M_0 > 0$ and $C' = -1$ for $T > 0$ we have

$$C = M_0' - T$$

[7] Department of Electrical Engineering, University of Wisconsin, Madison, Wis.

[8] A copy can be obtained by request to the Director, Engineering Experiment Station, Mechanical Engineering Building, University of Wisconsin, Madison, Wis.

$$M' = M_0' - \tau - T$$

$$M = M_0 + (M_0' - \tau)\,T - \frac{1}{2}\,T^2$$

Now $M = 0$ when

$$T = (M_0' - \tau) + \sqrt{[(M_0' - \tau)^2 + 2M_0]}$$

At this value of T

$$C = \tau - \sqrt{[(M_0' - \tau)^2 + 2M_0]}$$

For the next phase we keep $M \equiv 0$ by letting

$$\tau C' + C = 0 \qquad [141]$$

whence

$$C = \{\tau - \sqrt{[(M_0' - \tau)^2 + 2M_0]}\}e^{-T/\tau}$$

where we have taken $T = 0$ at the start of this phase. The initial C' is then

$$\left\{\frac{\sqrt{[(M_0' - \tau)^2 + 2M_0]}}{\tau} - 1\right\}$$

The quantity C' can attain this value only if

$$|\tau - \sqrt{[(M_0' - \tau)^2 + 2M_0]}| \leq \tau$$

This condition is not satisfied if

$$\sqrt{[(M_0' - \tau)^2 + 2M_0]} > 2\tau \qquad [142]$$

If $M_0 < 0$ the Inequality [142] is replaced by

$$\sqrt{[(M_0' + \tau)^2 - 2M_0]} > 2\tau \qquad [143]$$

Thus C can coast to equilibrium according to Equation [141] only if $|M_0|$ and $|M_0'|$ are sufficiently small.

Thus *even though the ideal transients are not obtained by letting the servo travel at only full or zero speed, there exist control functions* depending on M and C only that yield the ideal transients. In fact, these transients are obtained by letting

$$C' = -K\Sigma$$

where K is very large (infinite in the ideal case) and

$$\Sigma = M$$

for $M \neq 0$ while

$$\Sigma = \frac{C}{K\tau}$$

for $M = 0$, provided that $\tau \neq 0$.

That the author's enthusiasm for the nonlinear approach of the paper is colored by his work on governors is no doubt correct. Customers normally test prime-mover governors by making sudden load rejections or acceptances. Such sudden changes often occur in practice as when a generating unit is separated from the line and concern about the resulting transients is justified. However, the general principles of the paper were verified on a Philbrick analog computer simulating controlled systems of a rather general nature. These studies showed that the use of the absquare is beneficial for a wide range of applications where the disturbances are not necessarily of the step variety.

Dr. Stout is correct in stating that the theory serves as a guide. The improvement that can be obtained in practice by the application of the theory of the present paper was most disappointing to the author, who completed the theory as far as he felt it had to be carried for engineering applications, before initiating the experimental work to verify it. Improvement in maximum overswings of 2:1 for major disturbances without changing the quality of the transients is often about the best that can be attained where the application of the theory is indicated. In comparing linear and nonlinear controls one must be careful to employ the best control with a linear control function involving as many derivatives as is practical (up to the second) and arbitrary coefficients and subject to arbitrary discontinuities.

Professor Higgins is correct in his statement that the paper concerns on-off servomechanisms. The paper actually treats servomechanisms in general where the rate of change of the controlling variable is bounded, and this rate can be made to assume arbitrarily any value between its minimum and maximum. The paper is primarily concerned with transients that can be made optimum by operating a servomechanism as if it were of the on-off variety. Interest in the field of the paper is indicated by the other papers on the subject that are listed in Professor Higgins' superb bibliography on nonlinear control systems.

On the Analysis of Linear and Nonlinear Systems

By MARVIN SHINBROT,[1] MOFFETT FIELD, CALIF.

A general theory of a certain class of commonly used methods for the analysis of linear systems from measured response data will be described. It will then be shown that, when viewed from this general point of vantage, all of these linear techniques can be extended in a natural way to apply to nonlinear systems. In addition, through use of the general theory, a new method, possessing certain advantages over those used previously, will be derived. Finally, the effectiveness of the new method will be illustrated by several examples.

Introduction

THIS paper will be concerned with what has been called the "inverse" problem of system analysis, i.e., with the problem of determining the differential equations governing the behavior of a system, given a time history of the response of the system to some input. This problem has been of interest to aerodynamicists for some years, and the National Advisory Committee for Aeronautics has published several reports on the subject. Some of this work can be found in references (1 to 4).[2] Recently, workers in other fields appear to have become interested in the problem (see, e.g., reference 5).

Only two of the preceding references deal with nonlinear systems. The first of these (4) presents a method which, as mentioned in the report itself, is not only time-consuming but relatively inaccurate. Reference (5), which also concerns itself with nonlinear systems, applies only to those which are "slightly" so.

Last year, the National Advisory Committee for Aeronautics published (6). In this report, an entire class of methods was presented having no theoretical restrictions (such as slight nonlinearities, and the like) and with accuracy limited only by the accuracy of the experimental data themselves. Furthermore, the time required to apply one of the methods is no greater than that required by a method which has been used for years for the analysis of linear systems. The methods were obtained as a generalization of certain known linear techniques which were called "equations-of-motion" methods in (7). The present paper is devoted to a somewhat heuristic account of the theory and methods of (6).

The paper begins with a brief description of three of the most widely known equations-of-motion methods. This is presented in order to show how a general theory of such methods can be constructed. It is then shown how this general theory can be used to extend the methods to apply to nonlinear systems. Next, a new method, superior in certain respects to methods used heretofore, is discussed briefly. Two examples are displayed to indicate the efficacy of the new method.

In all that follows it will be assumed that there is only one equation of motion of the system and that it is of the second order. For more general situations, see reference (6).

Some Equations-of-Motion Methods

In reference (7), all of the known methods for linear systems were classified, somewhat unharmoniously, either as "response-curve-fitting" methods or as "equations-of-motion" methods. Since the methods of the former class all involve the assumption that the equations of motion may be solved explicitly in terms of the parameters of the system, clearly it would not be possible to extend them to apply generally to nonlinear systems. (Milliken would have classed the method of reference 5 as a response-curve-fitting method; hence the restriction in that method to slight nonlinearities.) As a consequence, the methods of (6) were generalizations of the equations-of-motion methods.

We shall begin with a discussion of three of these methods, as they apply to a linear system. It is assumed that records of two variables $x(t)$ and $F(t)$ are available. From these, it is desired to evaluate three constants b, k, and C, such that the given data satisfy the equation

$$\ddot{x} + b\dot{x} + kx = CF(t) \quad \ldots\ldots\ldots\ldots\ldots [1]$$

as closely as possible.

The first of the techniques to be discussed was given the name of the "derivative" method by Greenberg in (1). It consists in considering Equation [1] for fixed t as a linear equation in b, k, and C, instead of as an equation for x. Allowing t to vary results in a set of such equations which can be solved by least squares for the desired constants. This process gives the equations

$$\left.\begin{aligned} b\int_0^\infty \dot{x}^2\,dt + k\int_0^\infty \dot{x}x\,dt - C\int_0^\infty \dot{x}F\,dt &= -\int_0^\infty \dot{x}\ddot{x}\,dt \\ b\int_0^\infty x\dot{x}\,dt + k\int_0^\infty x^2\,dt - C\int_0^\infty xF\,dt &= -\int_0^\infty x\dot{x}\,dt \\ -b\int_0^\infty F\dot{x}\,dt - k\int_0^\infty Fx\,dt + C\int_0^\infty F^2\,dt &= \int_0^\infty F\ddot{x}\,dt \end{aligned}\right\} \ldots [2]$$

From the given record of $x(t)$, one can find $\dot{x}(t)$ and $\ddot{x}(t)$, calculate the integrals which are the coefficients in Equations [2], and then solve Equations [2] for b, k, and C. It should be noted that Equations [2] are obtained from Equation [1] by multiplying the latter by the functions $\dot{x}(t)$, $x(t)$, and $-F(t)$, respectively, and then integrating the results.

The "Laplace transform" method arises when one takes the Laplace transform of both sides of Equation [1]. Supposing for simplicity that $F(0) = x(0) = \dot{x}(0)$ (this is not essential to the method), the transformed equation is

$$bs\xi(s) + k\xi(s) - C\varphi(s) = -s^2\xi(s) \quad \ldots\ldots\ldots [3]$$

where $\xi(s)$ and $\varphi(s)$ are the transforms of $x(t)$ and $F(t)$, respectively. These transforms can be computed for several values of s

---

[1] Aeronautical Research Scientist, Ames Aeronautical Laboratory, National Advisory Committee for Aeronautics.

[2] Numbers in parentheses refer to the Bibliography at the end of the paper.

Contributed by the Instruments and Regulators Division of The American Society of Mechanical Engineers and presented at the ASME-AIEE Conference on Nonlinear Control Systems, Princeton, N. J., March 26–28, 1956.

Note: Statements and opinions advanced in papers are to be understood as individual expressions of their authors and not those of the Society. Manuscript received at ASME Headquarters, January 2, 1956. Paper No. 56—IRD-2.

and the resulting equations solved by least squares for b, k, and C. Note that Equation [3] is equivalent to

$$b\int_0^\infty e^{-st}\ddot{x}(t)dt + k\int_0^\infty e^{-st}x(t)dt - C\int_0^\infty e^{-st}F(t)dt = -\int_0^\infty e^{-st}\dot{x}(t)dt \quad [4]$$

since integrating by parts the appropriate integrals of Equation [4] gives

$$bs\int_0^\infty e^{-st}x(t)dt + k\int_0^\infty e^{-st}x(t)dt - C\int_0^\infty e^{-st}F(t)dt = -s^2\int_0^\infty e^{-st}x(t)dt$$

which is just Equation [3] with the functions $\xi(s)$ and $\varphi(s)$ replaced by their definitions in terms of the functions $x(t)$ and $F(t)$. Looking at Equation [4], it can be seen that Equation [3] is obtained by multiplying Equation [1] by e^{-st} and integrating. In the Laplace transform method, these integrals are computed for a certain number of values of s and the resulting Equations [3] solved by least squares for the desired constants.

The last method to be discussed is the so-called "Fourier transform" method. It is essentially the same as the Laplace transform method except that the Fourier transform replaces the Laplace transform.

For more details of these methods, see reference (1).

General Theory for Linear Systems

In order for the general theory to evolve, it is helpful to forget the ideas behind each of the methods described in the preceding section and to keep firmly in mind *the computational procedure which leads to each method.* Each of the methods considered can be described in these terms as follows:

The derivative method arises when

(a) Equation [1] is multiplied by the three functions $\dot{x}(t)$, $x(t)$, and $-F(t)$ one at a time.

(b) The resulting equations are integrated from zero to infinity.

(c) The equations to which step (b) leads are solved for b, k, and C.

We say the Laplace transform method is being applied when

(a) Equation [1] is multiplied by $N(\geq 3)$ functions $e^{-s_n t}$.

(b) The resulting equations are integrated from zero to infinity.

(c) The equations to which step (b) leads are solved (by least squares since N is usually chosen greater than 3) for b, k, and C.

The Fourier transform method results when, in the description of the Laplace transform method, the functions $e^{-s_n t}$ are replaced by $e^{-i\omega_n t}$ (or, alternatively, by $\cos \omega_n t$ and $\sin \omega_n t$).

That there is a general feature underlying the three methods which have been discussed is now clear. In each of them the same steps are performed, multiplication, integration, and solution of the resulting equations. Quite generally, then, one can develop equations-of-motion methods by selecting a set of functions $y_n(t)$ (called the "method functions"), $n = 1, \ldots, N$ for some $N \geq 3$, and then

(a) Multiplying Equation [1] by each of these functions.

(b) Integrating the resulting equations.

(c) Solving these new equations for the desired parameters.

In the three methods described, the integration of step (b) proceeds over the interval $(0, \infty)$; as will be seen, this is not essential and, indeed, the fact that it is inessential can be used noticeably to improve existing methods. To avoid some complications initially, we shall continue to integrate over the infinite interval; this restriction will be removed subsequently.

It should be noted that reference to a difference which exists in step (b) between the derivative method and the other two methods has been suppressed in order to get the ideas across better. This difference lies in the fact that in step (b) of the Laplace and Fourier transform methods, an integration by parts is performed. We shall discuss this again in the sequel.

The process described in the preceding paragraph leads, after step (b), to N equations of the form

$$b\int_0^\infty y_n\dot{x}\,dt + k\int_0^\infty y_n x\,dt - C\int_0^\infty y_n F\,dt = -\int_0^\infty y_n\ddot{x}\,dt \quad [5]$$

It is possible that functions $y_n(t)$ depend on the experimentally determined functions $x(t)$ and $F(t)$; this is the case for the derivative method, in which $y_1(t) = \dot{x}(t)$, $y_2(t) = x(t)$, $y_3(t) = -F(t)$. If this is so, Equations [5] can be considered as N equations which are to be solved (by least squares if $N > 3$) for the desired parameters. Of course, this process requires the (inherently inaccurate) calculation of the derivatives $\dot{x}(t)$ and $\ddot{x}(t)$ from the experimental data. If it is desired to eliminate such a step, the $y_n(t)$ must be chosen as explicitly independent of $x(t)$ and $F(t)$, as in the case of the Laplace or the Fourier transform methods. If this is the case, the following formulas, obtained by integration by parts, are used

$$\int_0^\infty y_n\dot{x}\,dt = -y_n(0)x(0) - \int_0^\infty \dot{y}_n x\,dt$$

$$\int_0^\infty y_n\ddot{x}\,dt = -y_n(0)\dot{x}(0) + \dot{y}_n(0)x(0) + \int_0^\infty \ddot{y}_n x\,dt$$

Substitution into Equations [5] gives

$$-b\left[y_n(0)x(0) + \int_0^\infty \dot{y}_n x\,dt\right] + k\int_0^\infty y_n x\,dt - C\int_0^\infty y_n F\,dt = y_n(0)\dot{x}(0) - \dot{y}_n(0)x(0) - \int_0^\infty \ddot{y}_n x\,dt, \quad n = 1, \ldots, N \quad [6]$$

Equations [6] are N equations in b, k, and C. If $N \geq 3$, they can be solved by least squares for those parameters.

Generalization to Nonlinear Systems

We have indicated how a general theory of equations-of-motion methods for linear systems is developed. We now proceed to consider nonlinear systems, in particular systems subject to the equation

$$\ddot{x} + f(x)\dot{x} + g(x) = 0 \quad [7]$$

It should be said that it certainly is possible to treat more general situations, in particular situations in which there is a forcing term on the right side of the equation, but Equation [7] suits admirably the needs of exposition.

In order to apply the method, $f(x)$ and $g(x)$ are approximated by polynomials. Although more complicated polynomials can be considered, we shall restrict ourselves here to the case where

$$\left.\begin{aligned} f(x) &= b_0 + b_1 x + b_2 x^2 \\ g(x) &= x(k_0 + k_1 x + k_2 x^2) \end{aligned}\right\} \quad [8]$$

There are no conceptual difficulties in the generalization to higher-order polynomials. The problem can now be stated as follows: Given an experimental record of $x(t)$, determine six constants b_i and k_i such that $x(t)$ satisfies Equation [7] as closely as possible, with f and g being given by Equation [8].

Manifestly, it is possible to generalize the method of the preceding section to solve this problem. Select N method functions $y_n(t)$. Multiply Equation [7] by these functions and integrate.

Integration by parts then gives

$$-b_0\left[y_n(0)x(0) + \int_0^\infty \dot{y}_n x\,dt\right] - \frac{1}{2}\,b_1\left[y_n(0)x^2(0) + \int_0^\infty \dot{y}_n x^2\,dt\right] - \frac{1}{3}\,b_2\left[y_n(0)x^3(0) + \int_0^\infty \dot{y}_n x^3\,dt\right] + k_0\int_0^\infty y_n x\,dt + k_1\int_0^\infty y_n x^2\,dt + k_2\int_0^\infty y_n x^3\,dt = y_n(0)\dot{x}(0) - \dot{y}_n(0)x(0) - \int_0^\infty \ddot{y}_n x\,dt, \quad n = 1, \ldots, N \ldots [9]$$

The integrals occurring in Equations [9] can be evaluated. Following this, Equations [9] can be solved for the desired constants.

Choice of the Method Functions

To this point in the discussion, the method functions $y_n(t)$ have been to a great extent arbitrary, having to satisfy only certain weak smoothness conditions so that the integrations by parts which led to Equations [6] and [9] were allowable. In this section, a few comments will be made on how this arbitrary character of the method functions can be utilized to devise new techniques, superior in point of accuracy to methods heretofore used.

With the exception of the derivative method, to which the following remarks do not apply (but which has its own troubles), any equations-of-motion method not carefully selected will be prone to certain unpleasant features. Upon inspection of Equation [6] or [9], the first thing which strikes the eye is the peculiar role played by the initial point $t = 0$. It is clear that a small change in the initial values can produce a sizable change in the calculated values of parameters while this is true of no other point. This feature is clearly undesirable. The cure can be seen from either of Equations [6] or [9] to be to choose the method functions such that

$$y_n(0) = \dot{y}_n(0) = 0, \quad n = 1, \ldots, N \ldots [10]$$

Choosing method functions with this property eliminates the heavy dependence on the initial conditions[3] and reduces Equation [9] to the following

$$-b_0\int_0^\infty \dot{y}_n x\,dt - \frac{1}{2}\,b_1\int_0^\infty \dot{y}_n x^2\,dt - \frac{1}{3}\,b_2\int_0^\infty \dot{y}_n x^3\,dt + k_0\int_0^\infty y_n x\,dt + k_1\int_0^\infty y_n x^2\,dt + k_2\int_0^\infty y_n x^3\,dt = -\int_0^\infty \ddot{y}_n x\,dt, \quad n = 1, \ldots, N \ldots [11]$$

Another objection to the methods discussed so far is the infinite interval of integration which occurs in Equation [11]. Naturally, no experimental data can ever be obtained over an infinite interval of time. Consequently, it would be nice if the infinite upper limit on the integrals occurring in Equation [11] could be replaced by a finite limit. Suppose the data available to be T seconds long. We cannot blindly replace ∞ by T in Equation [11], for the by-parts integrations which led to Equation [11] would give rise to terms, analogous to the terms eliminated by the Condition [10], depending on the final conditions when $t = T$. These terms cause the end conditions to be weighted unduly. They can be eliminated by adducing to Conditions [10] conditions of the form

$$y_n(T) = \dot{y}_n(T) = 0, \quad n = 1, \ldots, N \ldots [12]$$

If Conditions [10] and [12] are assumed, the basic Equation [11] becomes

$$-b_0\int_0^T \dot{y}_n x\,dt - \frac{1}{2}\,b_1\int_0^T \dot{y}_n x^2\,dt - \frac{1}{3}\,b_2\int_0^T \dot{y}_n x^3\,dt + k_0\int_0^T y_n x\,dt + k_1\int_0^T y_n x^2 dt + k_2\int_0^T y_n x^3\,dt = -\int_0^T \ddot{y}_n x\,dt, \quad n = 1, \ldots, N \ldots [13]$$

The method functions are still quite arbitrary; they only must satisfy Equations [10] and [12] in order that [13] follow. In order that all the data be given equal weight, the further condition that the method functions do not consistently approach zero somewhere may be imposed. (This eliminates the Laplace transform method and, for stable systems in which the output approaches zero, the derivative method.) In (6), method functions of the form

$$y_n = \sin^2 \omega_n t \ldots [14]$$

were considered. Naturally, these functions, while satisfying Equation [10], will not in general satisfy [12] unless the frequencies ω_n are properly chosen. If the ω_n are chosen according to the rule

$$\omega_n = \frac{n\pi}{T} \ldots [15]$$

Conditions [10] and [12] will both be satisfied. This choice of frequencies leads to an elegant method which gives satisfactory results in certain cases. On the other hand, the difference (π/T) between two successive frequencies is too large to define the "frequency response" (to use loosely the terminology of the Fourier transform) of some examples adequately unless T is quite large. Hence, in (6), more frequencies were inserted between the Frequencies [15]. In order that [12] continue to be satisfied, the method functions were chosen of the form of Equation [14], except that after the last zero of [14] in the interval $0 \leq t \leq T$, the function was cut off, so that some of the method functions were identically zero over a small part of the basic interval. For further details of this method, the reader is referred to (6). The efficacy of this choice of method functions is illustrated by the examples to follow.

Examples

Two examples will be given. The first is that of a slightly nonlinear hardening spring and the second that of the van der Pol oscillator.

For the first example the equation which it is assumed the data satisfy is

$$\ddot{x} + b\dot{x} + g(x) = 0 \ldots [16]$$

For the example under consideration, b was chosen equal to 2 and

[3] The Condition [10] would be replaced by the condition

$$y_n^{(k)}(0) = 0, \quad k = 0,1,\ldots, K-1$$

if K is the highest derivative occurring in the basic equation assumed to describe the motion.

$g(x)$ was chosen as the solid curve in Fig. 1.[4] The "measured" data were manufactured by solving Equation [16] on a REAC. These data are displayed in Fig. 2. The method described was then applied to these data, assuming the form

$$g(x) = x(k_0 + k_1x + k_2x^2) \quad [17]$$

for g. This procedure gave the results

$$b = 1.95 \qquad k_1 = -30.5$$
$$k_0 = 50.4 \qquad k_2 = 806$$

Since the starting value of b was 2, we see that b has been found with an error of 2.5 per cent. The values of k_0, k_1, k_2 were substituted into Equation [17] and the resulting function was plotted as the dotted curve in Fig. 1. It can be seen that the error at the least accurate point is less than 3 per cent.

For the second example, the equation

$$\ddot{x} + 10(x^2 - 1)\dot{x} + x = 0$$

was solved on the REAC giving the "data" of Fig. 3. It was pre-

[4] A multiplicative factor was omitted when Fig. 1 was plotted. In order to correct the figure, the vertical scale should be multiplied by —0.01. Note that the minus sign on this factor effectively inverts the figure, thus making the system stable, which is as it should be.

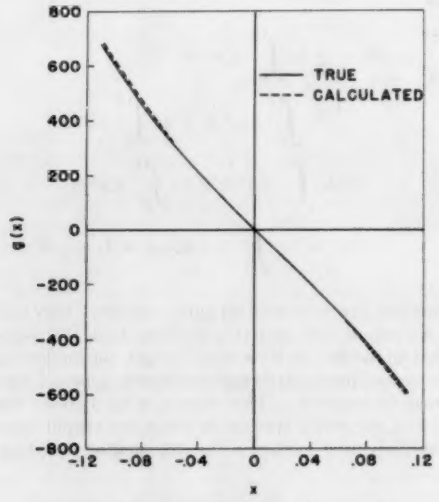


Fig. 1 Nonlinear Spring of Example 1

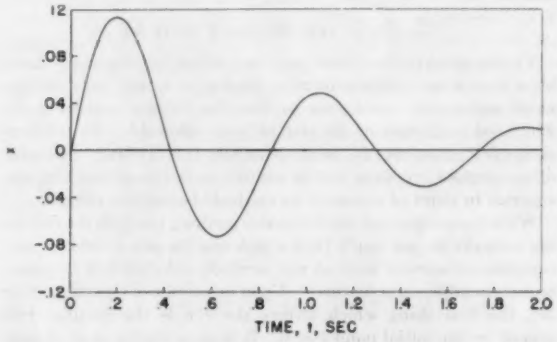


Fig. 2 Data for Example 1

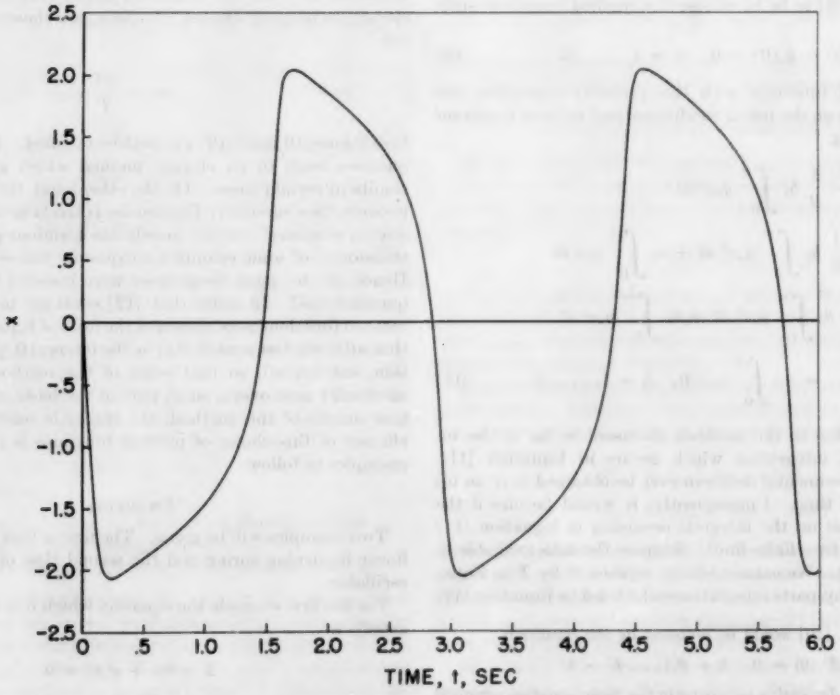


Fig. 3 Data for Example 2

tended that no more was known about the equation satisfied by the data than that it was of the form

$$\ddot{x} + (b_0 + b_1 x + b_2 x^2)\dot{x} + kx = 0$$

and the method was applied to this equation. This gave the results

$$b_0 = -10.0 \qquad b_2 = 9.42$$

$$b_1 = -0.14 \qquad k = 0.96$$

which clearly represents a good fit to the data.

Concluding Remarks

A general theory of the so-called equations-of-motion methods for the analysis of dynamical systems has been presented. It has been shown that, when looked at from a new point of view, all such methods can be generalized so as to apply to linear and nonlinear systems alike. Use of this theory also has shown how new methods can be developed to satisfy the requirements of particular problems.

One new method has been described in some detail. In certain cases, it reduces to the well-known Fourier transform method but in all cases has certain advantages over this latter method and over methods heretofore used. Its superiority is based on two facts: (*a*) There is the heavy dependence on the initial conditions which occurs when using most of the previously known equations-of-motion methods; this dependence is entirely eliminated in the new method. (*b*) The fact that most of the methods used to this time demand an infinitely long record for their rigorous application; this demand is not made by the new method. Finally, it also might be mentioned that the time of application of the method is no greater than that for existing methods and that the method is well suited to machine computation.

BIBLIOGRAPHY

1 "A Survey of Methods for Determining Stability Parameters of an Airplane From Dynamic Flight Measurements," by Harry Greenberg, NACA TN 2340, 1951.

2 "A Least Squares Curve Fitting Method With Applications to the Calculation of Stability Coefficients From Transient Response Data," by Marvin Shinbrot, NACA TN 2341, 1951.

3 "A Description and a Comparison of Certain Nonlinear Curve Fitting Techniques, With Applications to the Analysis of Transient-Response Data," by Marvin Shinbrot, NACA TN 2622, 1952.

4 "Techniques for Calculating Parameters of Nonlinear Dynamic Systems From Response Data," by B. R. Briggs and A. L. Jones, NACA TN 2977, 1953.

5 "The Attenuation of Damped Free Vibrations and the Derivation of the Damping Law From Recorded Data," by K. Klotter, Stanford University Division of Engineering Mechanics, Contract N6-ONR-251, Technical Report 23, November 1, 1953.

6 "On the Analysis of Linear and Nonlinear Dynamical Systems From Transient-Response Data," by Marvin Shinbrot, NACA TN 3288, 1954.

7 "Dynamic Stability and Control Research," by W. F. Milliken, Jr., presented at Third International Joint Conference of RAS-IAS, Brighton, England, September 3–14, 1951, Cornell Aero. Laboratory, Inc., CAL-39; also issued in Anglo-American Aero. Conf. Report, 1952, pp. 447–524.

Discussion

E. S. Smith.[5] The author's method for handling nonlinearities appears to be new and of wide usefulness. However, the method is of such generality that a short application to a typical control is needed in the closure for the paper to be of most use to its readers.

[5] Ordnance Engineer, Ballistic Research Laboratories, Aberdeen Proving Ground, Md. Fellow ASME.

Author's Closure

In response to Mr. Smith's suggestion, the first example given in the paper will be worked out in greater detail. The data of Fig. 2 are given in Table 1.

TABLE 1

t	$x(t)$	t	$x(t)$	t	$x(t)$	t	$x(t)$
0	0	0.50	−0.0413	1.00	0.0445	1.50	−0.0303
0.05	0.0455	0.55	−0.0620	1.05	0.0480	1.55	−0.0285
0.10	0.0833	0.60	−0.0720	1.10	0.0460	1.60	−0.0232
0.15	0.1070	0.65	−0.0718	1.15	0.0377	1.65	−0.0152
0.20	0.1140	0.70	−0.0612	1.20	0.0265	1.70	−0.0065
0.25	0.1053	0.75	−0.0448	1.25	0.0125	1.75	0.0025
0.30	0.0835	0.80	−0.0233	1.30	−0.0010	1.80	0.0100
0.35	0.0530	0.85	−0.0015	1.35	−0.0132	1.85	0.0158
0.40	0.0188	0.90	0.0183	1.40	−0.0233	1.90	0.0188
0.45	−0.0148	0.95	0.0333	1.45	−0.0290	1.95	0.0195
						2.00	0.0180

In accordance with Equation [14] the method functions were chosen of the form

$$y_n = \sin^2 \omega_n t \qquad [18]$$

over the first part of the interval, and as zero over the part of the interval after the last zero of y_n in $0 \leq t \leq 2$, as was discussed following Equation [15].

TABLE 2

n	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
ω_n	$\frac{\pi}{2}$	$\frac{10\pi}{13}$	π	$\frac{5\pi}{4}$	$\frac{3\pi}{2}$	$\frac{30\pi}{17}$	2π	$\frac{20\pi}{9}$	$\frac{5\pi}{2}$	$\frac{25\pi}{9}$	3π	$\frac{10\pi}{3}$	$\frac{7\pi}{2}$	$\frac{70\pi}{19}$	4π

The frequencies used in Equation [18] are given in Table 2. The frequency ω_1 is not entered in the table since $y_1 \equiv 0$ and so offers no information.

For this example, Equation [13] becomes

$$-b\int_0^T \dot{y}_n x\,dt + k_0\int_0^T y_n x\,dt + k_1\int_0^T y_n x^2\,dt + k_2\int_0^T y_n x^3\,dt = -\int_0^T \ddot{y}_n x\,dt,\ n = 2, \ldots, 16 \ldots [19]$$

From Table 1, the quantities x^2 and x^3 can be found, and then the integrals occurring in Equation [19] can be computed.[6] This computation gives the results displayed in Table 3.

Incidentally, it can be seen that the entries in Table 3 have from four to six significant figures, while the original data of Table 1 have fewer. Naturally, then, not all the figures of Table 3 have meaning, but they have been carried through the computation and roundoff is only performed at the end.

With parenthesized numbers referring to columns in Table 3, it can be seen that Equations [19] are equivalent to

$$(2)b + (3)k_0 + (4)k_1 + (5)k_2 = (6),\ n = 2, \ldots, 16$$

Hence, if these equations are to be solved by least squares, one must solve the Equations

[6] Simpson's rule of course can be used for this computation, but the method of L. N. G. Filon is better suited to it. For this method, see "On a Quadrature Formula for Trigonometric Integrals," by L. N. G. Filon, Proceedings of the Royal Society of Edinburgh, vol. 19, 1928–1929, pp. 38–47; "Integral Transforms in Mathematical Physics," by C. J. Tranter, John Wiley & Sons, Inc., New York, N. Y., 1951. The method is also described in an appendix to reference (6) of the present paper where tables of integration coefficients are given.

TABLE 3

(1)	(2)	(3)	(4)	(5)	(6)
n	$-\int \ddot{y}_n x\,dt$	$\int y_n x\,dt$	$\int y_n x^2\,dt$	$\int y_n x^3\,dt$	$-\int \ddot{y}_n x\,dt$
2	−0.007550	−0.001075	0.0016715	−0.00000265	−0.107630
3	−0.025990	−0.006730	0.0018970	0.00000025	−0.442175
4	−0.124265	−0.009360	0.0021645	0.00003525	−0.757450
5	−0.201610	0.015110	0.0022380	0.00011815	0.392470
6	−0.049340	0.022210	0.0024005	0.00017315	1.099785
7	−0.027140	0.013860	0.0024895	0.00018030	0.721525
8	−0.000660	0.015030	0.0028400	0.00017660	0.823325
9	0	0.011240	0.0033385	0.00017030	0.608745
10	0.005105	0.012640	0.0032335	0.00016535	0.693960
11	0.002965	0.009920	0.0026770	0.00016080	0.552885
12	0.007775	0.011685	0.0025320	0.00015970	0.659540
13	0.006545	0.009275	0.0023670	0.00014355	0.514865
14	0.010500	0.011040	0.0023815	0.00012490	0.586380
15	0.010765	0.009805	0.0023400	0.00010480	0.500355
16	0.007790	0.010705	0.0023440	0.00009305	0.554280

$\Sigma(2) \times (2) = 0.06041725$ $\quad \Sigma(3) \times (5) = 0.0000226548$

$\Sigma(2) \times (3) = -0.00263175$ $\quad \Sigma(3) \times (6) = 0.113807$

$\Sigma(2) \times (4) = -0.000841900$ $\quad \Sigma(4) \times (4) = 0.0000935345$

$\Sigma(2) \times (5) = -0.0000350735$ $\quad \Sigma(4) \times (5) = 0.00000474140$

$\Sigma(2) \times (6) = -0.175448$ $\quad \Sigma(4) \times (6) = 0.0175061$

$\Sigma(3) \times (3) = 0.00221335$ $\quad \Sigma(5) \times (5) = 0.000000272438$

$\Sigma(3) \times (4) = 0.000360318$ $\quad \Sigma(5) \times (6) = 0.00114957$

$$\left.\begin{aligned}
&b\Sigma(2) \times (2) + k_0\Sigma(2) \times (3) + k_1\Sigma(2) \times (4) \\
&\qquad + k_2\Sigma(2) \times (5) = \Sigma(2) \times (6) \\
&b\Sigma(3) \times (2) + k_0\Sigma(3) \times (3) + k_1\Sigma(3) \times (4) \\
&\qquad + k_2\Sigma(3) \times (5) = \Sigma(3) \times (6) \\
&b\Sigma(4) \times (2) + k_0\Sigma(4) \times (3) + k_1\Sigma(4) \times (4) \\
&\qquad + k_2\Sigma(4) \times (5) = \Sigma(4) \times (6) \\
&b\Sigma(5) \times (2) + k_0\Sigma(5) \times (3) + k_1\Sigma(5) \times (4) \\
&\qquad + k_2\Sigma(5) \times (5) = \Sigma(5) \times (6)
\end{aligned}\right\} \ldots [20]$$

The values of the sums occurring here are given below Table 3. Substitution of these sums into the appropriate places in Equations [20] and solution of the resulting equations gives the values of the parameters quoted in the body of the paper.

More examples with even further detail are given in reference (6).

The author of (5) has asked that the following be mentioned as a more readily available source of his article.[7]

[7] "The Attenuation of Damped Free Vibrations and the Derivation of the Damping Law From Recorded Data," by K. Klotter, Proceedings of the Second U. S. National Congress of Applied Mechanics, 1954.

Physical and Mathematical Mechanisms of Instability in Nonlinear Automatic Control Systems

By R. E. KALMAN,[1] NEW YORK, N. Y.

This paper is a critical examination of the stability problem in automatic control systems containing nonlinear elements. An attempt is made to classify and isolate essentially different phenomena, and to illustrate each type by means of simplified but representative examples. Particular attention is paid to the effects of system parameters and system inputs in provoking or destroying instability. The phenomena discussed are so diverse that they defy any over-all conclusions. Still, it is hoped that the tentative classification adopted here will be of help in recognizing important and unimportant aspects of problems in nonlinear control-system design.

Glossary of Special Terms

Nonlinear Differential Equation. The set of relations

$$\frac{dx_i}{dt} \equiv \dot{x}_i = f_i(x_1, \ldots x_n) \quad (i = 1, \ldots n) \quad (*)$$

where f_i is some function of the variables $x_1, \ldots x_n$; f_i must be specified in each case.

Phase Space. An n-dimensional Euclidean space where the Cartesian co-ordinates of a point are the $x_1, \ldots x_n$, also called "phase-space variables." Physically, the $x_1, \ldots x_n$ determine n different stored energies in a lumped-parameter system. It is also said that the $x_1, \ldots x_n$ represent the "initial conditions" of the system. The phase space is the set of all possible initial conditions or "states" of a dynamic system.

Trajectory. A curve in the phase space which is specified in parametric form by a set of transient solutions of (*): $x_1(t), \ldots x_n(t)$; it starts at the point $x_1(t = 0), \ldots x_n(t = 0)$. The phrase, "the trajectories move" (or "the trajectories tend to . . . ") refers to the paths described by the trajectories as t changes monotonically (usually $t \to +\infty$).

Limiting Behavior. Description of motion of trajectories as $t \to +\infty$ or $t \to -\infty$.

Critical Point. A point in phase space where all the time derivatives $\dot{x}_1(t), \ldots \dot{x}_n(t)$ vanish simultaneously; sometimes called a "singular point."

Equilibrium Point = critical point—a point where a dynamic system is at rest.

Limit Cycle. A nonself-intersecting closed curve in the phase space which is a trajectory. It is "stable" if trajectories in its neighborhood tend to it as $t \to +\infty$; it is "unstable" if trajectories in its neighborhood tend to it as $t \to -\infty$. In the time domain the limit cycle corresponds to a periodic function.

Types of Critical Points. Critical points may be classified with the help of the characteristic roots of the linear differential equation with constant coefficients which govern the motion of trajectories near the critical point.

(*a*) *Node.* The characteristic roots are real and of the same sign.

(*b*) *Focus.* The characteristic roots are complex conjugate.

(*c*) *Saddle Point.* The characteristic roots are real and of opposite sign.

Note: This terminology applies only to critical points in two-dimensional phase space.

Stability of Critical Points. If the real part of each of the characteristic roots of the linear differential equation associated with the critical point is negative, the critical point is stable; if the real parts are all non-negative, the critical point is unstable. A saddle point is neither stable nor unstable.

The terminology defined in the foregoing is standard throughout the majority of the literature of nonlinear differential equations and nonlinear mechanics. For additional information consult references (1 to 4).[2]

1 Introduction

This paper is concerned with automatic control systems by which are meant servomechanisms, regulators, analog computers, gun-directors, autopilots, etc. Such systems may be described by the adjectives active, deterministic, dynamic. The class of systems treated in the paper is required to be, in addition, also time-invariant and lumped-parameter. Under such restrictions, the behavior of a system over time will be governed by an autonomous ordinary nonlinear differential equation.

The paper itself might be called a study of nonlinear differential equations. While correct, such a viewpoint would be very artificial, since all of the discussion which follows is motivated by practical engineering problems. To be sure, it will be unavoidable to deal with such problems in a somewhat abstract way, assuming the reader's familiarity with equipment used in the automatic control art.

A nonlinear system is defined by the fact that the principle of superposition does not apply. In other words, a linear combination of two solutions of a nonlinear differential equation is not, in general, a solution.

Further qualifications are necessary to arrive at the heart of the problem. As a consequence of the principle of superposition, the response of a linear system to a step of any magnitude can be obtained by merely scaling the unit step response. More loosely, it may be said that all step responses look alike. In a nonlinear system, the step responses will not, in general, be scalar multiples of one another, but it may happen that they still look alike. Whenever this is true, the effect of the nonlineari-

[1] Department of Electrical Engineering and Electronics Research Laboratories, Columbia University.

Contributed by the Instruments and Regulators Division of The American Society of Mechanical Engineers and presented at the ASME-AIEE Conference on Nonlinear Control Systems, Princeton, N. J., March 26–28, 1956.

Note: Statements and opinions advanced in papers are to be understood as individual expressions of their authors and not those of the Society. Manuscript received at ASME Headquarters, February 1, 1956. Paper No. 56—IRD-16

[2] Numbers in parentheses refer to references at end of paper

ties in the system is of minor importance practically and of limited interest theoretically. It also may happen, however, that the step responses will differ in some essential fashion; for instance, they may be nonoscillatory for a step input of sufficiently small magnitude and oscillatory for a step input of sufficiently large magnitude. (See Example 1 of reference 1.)

Going still further, nonlinear systems may be scrutinized according to what happens to the transient solutions as the time approaches infinity. Very crudely speaking, this is the notion of stability. It may happen that all transient responses tend to a unique equilibrium point in the phase space, regardless of initial conditions or the magnitude of the input. But it also may happen that not all, if any, transient solutions will tend to a unique equilibrium point.

If there are only two energy storage elements, i.e., the system obeys a second-order nonlinear differential equation, then classical results due to Poincaré and Bendixson (2, 3) state that the limiting behavior of unforced solutions (trajectories) of the differential equations as the time $t \to \infty$ has the following possible forms:

(a) The trajectories may tend to one or more stable equilibrium points.

(b) The system may oscillate continually, the trajectories tending to a limit cycle in the phase space.

(c) The trajectories may tend to infinity.

When the differential equation is of higher than second order, results similar to the foregoing are not yet available, but, in the author's opinion, there is little reason to expect surprises (4). Much more serious is the restriction that the foregoing results apply only to the autonomous behavior of a system; i.e., in the absence of an externally applied forcing function or input signal. Since most control systems depend very crucially on the inputs, this would seem to represent a serious difficulty. If, however, the input is of a simple type, such as a combination of a step and a ramp, it is usually still possible to carry out an analysis in such a fashion that the magnitudes of the steps and ramps become parameters of the system and affect its entire dynamic structure. This is illustrated in Example 1 and the same procedure is used without additional remarks in the other examples (see also references 1, 4).

Linear systems behave either according to (a) or according to (c) but not both. Nonlinear systems which behave similarly will be called *monostable*. Thus linear systems are always monostable. A monostable nonlinear system is topologically equivalent to a linear system; in other words, there exists some one-to-one bicontinuous transformation which deforms the trajectories of a monostable nonlinear system into those of a linear system. An example of a monostable nonlinear system is the equation: $\dddot{x} + a\ddot{x} + b\dot{x} + cx^3 = 0$ ($a, b, c > 0$).

If a nonlinear system is not monostable, all three types of limiting behavior may be encountered. For instance, van der Pol's equation $\ddot{x} + k(x^2 - 1)\dot{x} + x = 0$ is not monostable since it is known that all unforced solutions tend to a stable limit cycle when $k > 0$.

The fact (b) that sustained oscillations may be possible is one of the great discoveries of nonlinear mechanics. It is a physical phenomenon with which a control-system designer is well acquainted; he regards it usually as a nuisance, seldom as a friend. Case (c), on the other hand, is mathematical fiction indicating continual increase in energy storage which cannot take place in a physical universe. Realistically speaking, however, the existence of Case (c) is a tipoff that the equipment has a pronounced tendency to burn fuses or go up in flames—hardly to be ignored by the engineer.

The three cases listed are bothersome problems for the nonlinear control-system designer. His aim is to make the system behave in a uniform, predictable fashion—as much like a linear system as possible. (Sometimes his viewpoint may be different, for instance when building an oscillator. These cases will not be of interest here.) To wit, he has to make sure that there is only one stable equilibrium point, that no sustained oscillations are possible, and certainly that there cannot be explosively increasing energy storage. In other words, from his standpoint any limiting behavior other than a unique stable equilibrium point may be referred to as an "instability."

2 The Theme of the Paper

What are the physical and mathematical mechanisms leading to some form of instability, which are of interest in the design of current-day automatic control equipment?

Since the equilibrium points are readily calculated from the differential equation itself, avoiding more than one stable equilibrium point is a simple matter and does not merit further discussion. Undoubtedly the most important instability of interest in control systems are limit cycles; the circumstances which lead to the creation or destruction of limit cycles are to be studied in considerable detail. Solutions tending to infinity are also fairly readily avoidable and will receive much less attention.

The discussion will proceed by means of highly idealized, but typical, examples. In any realistic situation it is to be expected, of course, that none of the phenomena studied will occur in its purest form but is likely to interact and be more or less obscured in the over-all performance of the system.

The examples are grouped into several categories which, in the author's belief, represent basically different phenomena. The real usefulness of such a classification can of course be decided only after much further experience and study.

Most of the analysis will be carried out by means of the phase-plane technique (1 to 5); in particular, heavy use will be made of the ideas and terminology of the author's recent paper (1) in which organization of the phase-plane analysis of a wide class of systems is simplified and systematized.

3 Gain-Controlling Nonlinearities

A very commonly encountered situation is the nonlinear control system shown in Fig. 1. The system consists of a linear transfer function $G(s)$ which accounts for all the energy storage, and a nonlinearity $f(e)$ which does not depend on time. It is sometimes stated that $f(e)$ is a "static" nonlinearity; this terminology is misleading and confusing because nonlinearities can be specified independently of any energy storage—provided only that the over-all system is describable by a differential equation.[3]

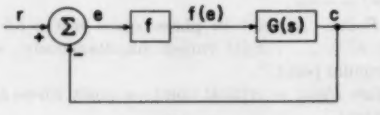


Fig. 1

3.1 *Sufficient Condition for Stability and Necessary Condition for Instability*

If $f(e)$ is a single-valued, continuous, differentiable function, its slope $f'(e)$ defines an incremental loop gain. By this is meant the following: As long as the transient stays in the neighborhood of $e = e_0$ in the phase space, the nonlinearity may be replaced by a constant $K = f'(e_0)$, and the behavior of trajectories near $e = e_0$ will be governed by a linear differential equation (with constant

[3] Separating nonlinearities from energy storages may lead to rather complex block diagrams.

coefficients) whose characteristic roots depend on K and $G(s)$. This self-evident and seemingly naive notion actually has important consequences:

(a) If $f(e)$ in Fig. 1 is replaced by constants K corresponding to all possible values of $f'(e)$, and it is found that the closed-loop system is stable for all such K, then it is intuitively clear that the system *must* be monostable; i.e., all transient solutions will converge to a unique, stable critical point.

(b) If the closed-loop system is unstable for some K and stable for other K, i.e., if the roots of the characteristic equation have both positive and negative real parts depending on K, then one would expect that the system *may* not be monostable.

Since K is a factor in the closed-loop gain of the system, it is logical to call $f(e)$ a *gain-controlling nonlinearity*. The usefulness of the concept is dependent, of course, on being able to determine quickly the closed-loop roots as a function of K. Fortunately, this problem has been basically answered by the recently developed *root-locus method* (5 to 7).

To prove the foregoing statements, one may proceed as follows: Suppose that the nonlinearity $f(e)$ is approximated by straight-line segments, as in Fig. 2. Each of these segments corresponds to a fixed value of K. The approximation may be thought to be equivalent to sampling the root locus; there must be a sufficiently large number of values of K (i.e., straight-line segments) to preserve the essential features of the root locus.

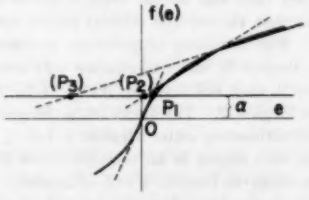


FIG. 2

To each straight-line segment there corresponds a region in phase space within which the transient solutions obey a linear differential equation. The characteristic roots of these linear differential equations are determined by the particular slope K of the straight-line segment. It is well known that all trajectories of an autonomous stable linear differential equation tend to a unique critical point in phase space as the time $t \to \infty$. Where is the critical point of a given region? Suppose the critical points of the nonlinear system satisfy $f(e) = \alpha$, where α is a real constant.[4] Then the value of e corresponding to the critical point of each phase-space region is determined by the intersection of the (extended) straight-line segments with the line $\alpha = \text{const}$, as shown in Fig. 2. Intersections lying on segments of the straight-line approximation of $f(e)$ define actual critical points of the system (P_1 in the figure). All other points may be called *virtual critical points*, since they lie outside the phase-space region to which they belong and therefore they can never be reached by their own trajectories. It will be convenient to use brackets to distinguish virtual critical points [(P_2), (P_3) in Fig. 2] from actual ones. Further details of this procedure may be found in an earlier paper by the author (1).

Now suppose that the nonlinearity f has been approximated with a suitably large number of straight-line segments. Suppose as in assumption (a), that the critical point of each phase-space region associated with the segments represents a stable linear system, regardless of whether the critical point is actual or virtual. Take an actual critical point. Find all trajectories belonging to the critical point and extend each trajectory backwards in time until the boundaries of the region are reached. If the adjacent region also had an actual critical point [which must be stable by assumption (a)], then the trajectories somewhere along the boundary must point in two different directions. This would contradict the fact that the solutions are unique when the nonlinearity is continuous. Hence all regions adjacent to a region with an actual critical point will have virtual critical points because of assumption (a). Continuing the same process, it follows that the system can have but one actual critical point.

To conclude that the system is monostable, it is also necessary to show that there are no limit cycles. This follows more easily from a physical argument. If there were a limit cycle, then the net energy lost in traveling around it in the phase space must be zero (cf. also Example 1). But under assumption (a), the system is everywhere dissipative and the net energy loss along any curve in the phase space is positive. Hence there cannot be any limit cycles.

The arguments are clearly not affected by the number of nonlinearities in the system. They depend crucially, however, on the *continuity* of the nonlinearities (cf. also Examples 4 to 6). The foregoing is summarized by the following:

Theorem. An n-th order ordinary nonlinear differential equation containing an arbitrary number of single-valued, continuous nonlinearities is monostable if the incremental linear differential equation at each point in the phase space is stable.

Proposition (b) is simply the converse statement of this theorem; i.e., the theorem implies the following:

Corollary. The differential equation just referred to fails to be monostable only if the incremental linear differential equation is unstable in some region in the phase space.

Observe that the theorem states a sufficient and the corollary a necessary condition. It is curious that this intuitively obvious and yet very general and useful result has apparently not been proven until now in the engineering literature.

Monostable systems are of no interest in this paper. Examples 1 to 3 which follow are nonlinear systems which are not monostable, but where the nonlinearities are continuous.

3.2 *Examples of Systems With Gain-Controlling Nonlinearities*

Example 1. As the first illustration, consider a control system such as shown in Fig. 1, assuming arbitrarily that

$$\left.\begin{aligned} G(s) &= (1 - s)/s(1 + s) \\ f(e) &= \text{saturation curve} \end{aligned}\right\} \quad \ldots\ldots\ldots\ldots [1]$$

In physical terms, one might think of the situation as involving the control of a system with integration and deadtime, where the effectiveness of the controlling variable $e(t)$ is greatly reduced beyond a certain linear range—one says that e "saturates." Such a system would have the transfer function $G(s) = \epsilon^{-2s}/s$; for convenience, however, ϵ^{-2s} is approximated with a lumped-parameter transfer function[5] $(1 - s)/(1 + s)$, which agrees with the first three terms of the Taylor expansion of ϵ^{-2s} about $s = 0$. The saturation curve is shown in Fig. 3; a saturation curve may be specified mathematically by

$$\begin{aligned} f(e) &= -f(-e) \\ f'(e) &= \text{a nonincreasing positive function of } |e| \\ \max f'(e) &= f'(0) \end{aligned}$$

[4] In general, e will be a linear combination of the phase-space co-ordinates; additional conditions (e.g., $\dot{e} = 0$, $\ddot{e} = 0$, etc.) are needed to fix the specific value of each of the phase-space co-ordinates at the critical point. But this is of no interest here, because the effect of the nonlinearity depends only on the value of e at the critical point.

[5] Actually, simple problems involving deadtime can be treated directly in the phase plane. See Eckman (8).

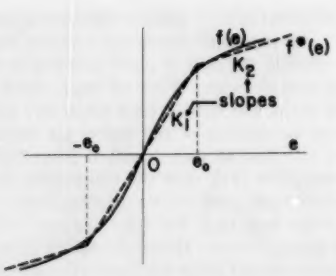


FIG. 3

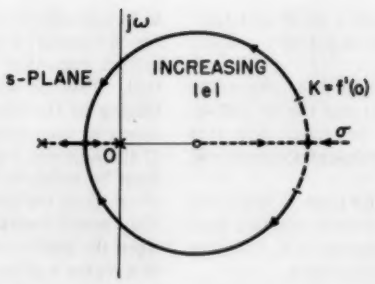


FIG. 4

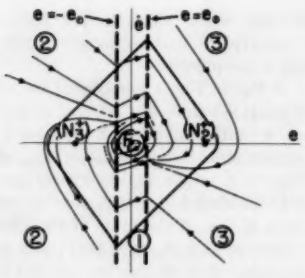


FIG. 5

A glance at the root-locus curve for the system (Fig. 4) shows that if $f'(0)$ is sufficiently large and if $f'(|e|)$ is nonincreasing, the root locus as a function of $|e|$ will be initially in the right half plane and later in the left half plane. This suggests that it is sufficient to take straight-line segments with only two different slopes, as shown in the figure, to approximate $f(e)$. The straight-line approximation is denoted by $f^*(e)$. If a more detailed study of trajectories is called for, then it may be desirable to take as many as four different slopes to account for the fact that the linear system corresponding to each straight-line segment in the approximation of $f(e)$ may have the following types of critical points: Unstable node, unstable focus, stable focus, stable node (arranged according to decreasing $f'(|e|)$ (1).

From Equation [1], the differential equations of the system are

$$\left.\begin{aligned} r(t) - c(t) &= e(t) \\ f^*(e(t)) &= y(t) \\ \ddot{c}(t) + \dot{c}(t) &= y(t) - \dot{y}(t) \end{aligned}\right\} \quad \ldots\ldots\ldots\ldots [2]$$

The dots denote differentiation with respect to time.

The problem of considering the effect of inputs on the dynamic structure of the system has already been pointed out in Section 1. Briefly, the procedure is as follows: The input is to have some simple functional form, which depends on a number of parameters. When the input function so specified is substituted in the differential equations of the system, the input parameters become system parameters. The procedure is successful only if the resulting differential equations are time invariant. This requires, in the present case, that the input consist of combinations of steps and ramps only. Using the constant r to denote the magnitude of the step component and $\dot{r}$ the magnitude of the ramp component of the input, the class of admissible inputs is given by

$$\left.\begin{aligned} r(t) &= r + \dot{r}t \quad t \geq 0 \\ &= 0 \qquad t < 0 \end{aligned}\right\} \quad \ldots\ldots\ldots\ldots [3]$$

Noting Equation [3], y and c are eliminated from Equations [2] bearing in mind that the derivative of $f^*(e)$ is discontinuous

$$\left.\begin{aligned} \frac{de}{dt} &= \dot{e} \\ \frac{d\dot{e}}{dt} &= -f^*(e) - (1 - K_1)\dot{e} + \dot{r} \quad \text{if } |e| \leq e_0 \\ &= -f^*(e) - (1 - K_2)\dot{e} + \dot{r} \quad \text{if } |e| > e_0 \end{aligned}\right\} \quad \ldots\ldots [4]$$

K_1 and K_2 are defined in Fig. 3. The critical points of Equations [4] are located at

$$f^*(e) = \dot{r} \quad \text{and} \quad \dot{e} = 0 \quad \ldots\ldots\ldots [5]$$

There are several interesting cases arising in connection with the system [4].

(i) Let $K_1 > 1$ and $K_2 < 1$; also let $r(t) \equiv 0$. Then the root locus will be in the right half plane for small $|e|$ and in the left half plane for large $|e|$. Also, Equations [4] satisfy the conditions of a well-known theorem of Levinson and Smith (9), and it follows at once that there exists a unique, stable limit cycle which represents the limiting behavior of all trajectories.

It is much more instructive, however, to employ the idea of actual and virtual critical points mentioned in Section 3.1 to explore the stability of the system. There will be always one actual and two virtual critical points; the former always unstable, the latter stable. It does not matter much whether the critical points are foci or nodes; for purposes of illustration it may be assumed that the actual critical point is an unstable focus, F_1^-, and that the virtual critical points are stable nodes (N_2^+), (N_3^+). The resulting trajectories are sketched in Fig. 5. If $e > e_0$, Region 3, the trajectories will tend to (N_3^+); if $e < -e_0$, Region 2, they will tend to (N_2^+). (N_2^+) is in Region 3 and (N_3^+) is in Region 2. It follows from the geometry that all trajectories will ultimately enter Region 1, $|e| \leq e_0$. Since the critical point in this region is an *unstable* focus F_1^-, the trajectories cannot remain in Region 1 but ultimately must leave it. Thus a situation has arisen where the trajectories cannot end up at a critical point since the only actual critical point is unstable; moreover, they cannot tend to infinity since the virtual critical points are stable. The only possible limiting behavior for the trajectories then is to tend to a stable limit cycle.

The argument used here may be sharpened by a device due originally to Poincaré: Since the system is stable for large $|e|$, there exists a very large closed curve in the phase plane such that all trajectories cross it toward the inside. Moreover, since the actual critical point of the system is unstable, there exists a very small closed curve in the neighborhood of the critical point, containing the critical point, such that all trajectories cross it toward the outside. This implies that there exists at least one stable limit cycle in the annular region between the two closed curves.[6] The two closed curves in question are also shown in Fig. 5; the large one is diamond-shaped, the small one is a circle. By examining the limiting behavior of the trajectories, it is apparent that the limit cycle (which is not sketched, to avoid congestion) has a shape roughly midway between that of the two closed curves. Unfortunately, the construction of such closed curves is a delicate analytical task and therefore definitely not recommended as a practical procedure. The concept of actual and virtual critical points, however, is very easy to handle and it can be extended readily to systems of higher than second order (4) although this cannot be discussed here in detail. See, however, Example 2, which follows.

Notice that the limit cycle lies in all three regions. This suggests a simple physical explanation of instability. The limit cycle exists if and only if the energy gained during the time the trajectory was in Region 1 (where the damping is negative) is

[6] Reference (2), p. 245.

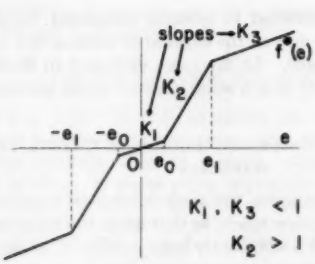


FIG. 6

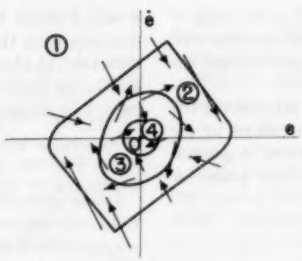


FIG. 7

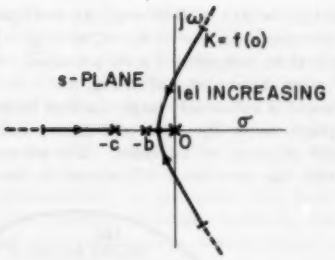


FIG. 8

precisely equal to the energy lost while the trajectory was in Regions 2 and 3 (where the damping is positive). This is as far as it is possible to go with intuitive physical explanations. To show, for instance, that the limit cycle is actually stable (or that it can exist at all—not a trivial problem as will be seen in Case ii), it is necessary to use arguments which are explicitly or implicitly the analog of the phase-plane picture Fig. 5.

(ii) The question arises as to what the effect of a nonzero input is. If the input is a step of any magnitude, then there is no effect whatever since the constant r simply does not appear in Equations [4]. This is because the control system discussed has zero steady-state error in response to step inputs (1, 4).

If the input is allowed to be a ramp, the situation will change because $\dot{r}$ appears as a parameter in Equations [4]. Now if the actual critical point as given by Equation [5] is still located at $|e| \leq e_0$, all the discussion in connection with Case (i) remains valid, although the phase-plane picture will become distorted. If the actual critical point is at $|e| > e_0$, the situation changes abruptly, because now the unstable actual critical point has become stable. The same "exchange of stabilities" will take place in regard to the virtual critical points.[7] If there exists a limit cycle it must enclose an actual critical point. If the critical point is stable, the limit cycle enclosing it must be necessarily unstable. Since there is no intuitive reason to expect an unstable limit cycle to occur, it seems that increasing $|\dot{r}|$ destroys the stable limit cycle. The rigorous analysis is as follows:

As $\dot{r}$ increases from 0, all critical points move to the right but the linear regions remain unchanged. As $f^{*-1}(\dot{r})$ approaches the upper breakpoint e_0 in the approximated saturation curve shown in Fig. 3, the distance between $(N_3{}^+)$ and $F_1{}^-$ decreases; in fact, $(N_3{}^+)$ moves from Region 2 into Region 1. By examining the trajectories belonging to Regions 1 and 3, it is easily seen that as $(N_3{}^+)$ and $F_1{}^-$ approach one another, the stable limit cycle becomes smaller and smaller until it coincides with the actual critical point when $\dot{r} = f^*(e_0)$. If we let $\dot{r} > f^*(e_0)$, the limit cycle disappears altogether and the critical points become $(F_1{}^-)$, $N_2{}^+$, and $(N_3{}^+)$, all being located in this order on the e-axis in Region 3. Thus the unstable actual critical point and the stable limit cycle coalesce and give rise to a stable critical point as $|\dot{r}|$ increases from zero. After Poincaré, phenomena of this type are called *bifurcation* (2, p. 73).

The dependence of stability of a nonlinear system on a parameter and, in particular, the creation and destruction of limit cycles is an important, profound, and relatively little explored question of the theory of nonlinear differential equations. A brilliant recent study by Duff (10) suggests that useful results can be obtained by quite simple means. In fact, the foregoing analysis is motivated by Duff's work, although his technique is not directly applicable here.

(iii) To explore further the problems raised in the last case, the nonlinearity is replaced by that shown in Fig. 6 where a straight-line approximation has been made already. The difference is that now the root locus begins in the left half plane, moves into the right-half plane, and finally returns to the left half plane as $|e|$ increases from zero, since $f'(|e|)$ is not monotonic. Three different values of $f'(|e|)$ are sufficient to sample the root locus (as was assumed in drawing Fig. 6), but if only two values had been taken, some aspect of stability would have been lost.

Assume again $r(t) \equiv 0$. The origin is a stable actual critical point and there are stable virtual critical points corresponding to K_3. But there are also unstable virtual critical points corresponding to K_2. Hence one may suspect that possibly not all trajectories will reach the origin. Should this be correct, it is possible to construct three nested closed curves as sketched in Fig. 7. If such a set of closed curves exists, then all trajectories leave Region 1 and tend to a stable limit cycle in Region 2. All trajectories entering Region 4 end up at the stable critical point at the origin. There exists an unstable limit cycle in Region 3 such that trajectories on its outside tend to the stable limit cycle and on its inside to the critical point at the origin. The existence of the largest and smallest such closed curves may be confidently assumed by virtue of arguments used in Case (i). But there are no *a priori* grounds for assuming the existence of a closed curve in the middle. This is strictly a quantitative question. If $K_2 \gg 1$ and $e_1/e_0 \gg 1$ (see Fig. 6), then the existence of such a closed curve appears plausible on physical grounds. As either one of the parameters K_2 or e_1/e_0 is decreased, a situation similar to Case (ii) may be expected to occur. The two limit cycles approach one another and at some stage, when $K_2 > 1$ and $e_1/e_0 > 1$ still hold (i.e., when there is still a pair of unstable virtual critical points), the limit cycles coincide. If the parameters are further decreased, the limit cycles disappear and, *a fortiori*, it will not be possible to construct a closed curve such that all trajectories cross it toward the outside.

The discussion of the effect of ramp input is quite similar to that of Case (ii) and may be omitted.

Example 1 has shown how the stability of a system may be influenced by the nonlinearity, keeping the transfer function unchanged. The converse situation is also interesting.

Example 2. The system is again as shown in Fig. 1. Let $f(e)$ be a saturation-type nonlinearity, Fig. 3. Consider two different open-loop transfer functions:

(i) $G(s) = A/s(s + b)(s + c)$ (A, b, c = positive constants). The corresponding root locus is sketched in Fig. 8. If A is sufficiently large, the root locus will start in the right half plane and move monotonically into the left half plane with increasing $|e|$ as indicated in the figure. Arguments similar to those used in connection with Case (i), Example 1, show that there exists a unique stable limit cycle surrounding an unstable critical point in the phase space. By making A sufficiently small, the critical point becomes stable and the limit cycle is destroyed.

[7] If $f(e)$ rather than its approximation $f^*(e)$ is used in Equation [5], the transition is not abrupt; otherwise the situation is the same.

(ii) $G(s) = A(s + d)(s + g)/s(s + b)(s + c)$ (A, b, c, d, g = positive constants, $0 < b < c \ll d < g$). In this case the root locus is more complicated and, provided $c \ll d$, it will have the qualitative shape sketched in Fig. 9.

When A is sufficiently large, the root locus will start in the left half plane, enter the right half plane, and then return to the left half plane as $|e|$ increases. The assumption $c \ll d$ guarantees that the root locus will remain in the right half plane for

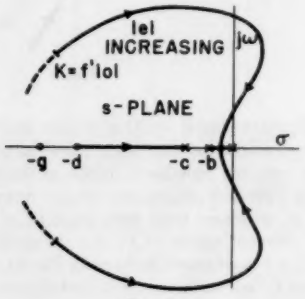


FIG. 9

a large range of $|e|$; arguments similar to those used in Case (iii) Example 1, show that there exists an unstable and a stable limit cycle in the phase space in addition to a stable critical point.

If the condition $c \ll d$ is not satisfied with a sufficient margin, then the root locus will not remain in the right half plane long enough and no limit cycle will exist. In other words, as the zero at $-d$, originally deeply in the left half plane, moves along the real axis toward the origin, the limit cycles in the phase space will move closer and closer together until they coincide and thereafter vanish. It should be noted also that Kochenburger's approximate stability criterion (11) implies the existence of a stable and an unstable limit cycle as long as the root locus crosses the imaginary axis twice. In reality, the preceding discussion shows that the limit cycles will vanish before the root locus leaves the right half plane entirely. The exact *sufficient* conditions for the existence of limit cycles would be extremely difficult to find and should best be left to computer studies in particular cases.

Finally, as A decreases, it may happen that the root locus starts in the right-half plane. Then the situation is the same as Case (i) discussed previously.

Example 3. In a recent paper, Lewis (12) suggests improving the performance of a conventional second-order servo by making the damping ratio small or even negative when the error is large. The equations of such a system in the $(e, \dot{e})$ phase plane are

$$\frac{de}{dt} = \dot{e}; \quad \frac{d\dot{e}}{dt} = -[2(b - a|e|)\dot{e} + e] \text{ (step inputs only)} \ldots [6]$$

Inspection of Equation [6] shows that trajectories which remain in the region $b - a|e| > 0$ tend to the unique critical point at the origin of the phase plane; but if a trajectory enters the region $b - a|e| < 0$ it may diverge to infinity, i.e., there may be "explosive" instability. This has been confirmed by means of an analog-computer study (13). Indeed, by detailed analysis of Equation [6], which is too complicated to be given here, it is possible to establish the existence of a unique, unstable limit cycle.

These illustrations show the usefulness of the root-locus concept in obtaining a quick feeling for possible instabilities in a control system. Recall, however, that the nonlinearities must be continuous. A much richer storehouse of phenomena is encountered in the next sections where discontinuous nonlinearities of various types are studied. The systems considered in Sections 4 and 5 could be converted to systems containing continuous nonlinearities, though only at the expense of working in a higher-dimensional phase space. In the cases discussed in Sections 6 and 7, the discontinuity is in a sense inherent in the problem and cannot be eliminated.

4 Systems Whose Governing Equations Change Discontinuously

In the preceding examples, the *same* differential equation was valid over the entire phase space, so that when the nonlinearities are approximated with a sufficiently large number of straight-line segments, the properties of the corresponding linear systems in adjacent regions of the phase space change in a continuous manner. One way of relaxing the continuity is by letting the system be governed by two entirely different linear differential equations in adjacent, nonoverlapping regions of the phase space. Probably the simplest example is provided by coulomb ("dry") friction.

Example 4. Consider a simple positional servomechanism, where the output element, in addition to inertia, is subject to combined coulomb and viscous friction. The torque T to the output element is supplied by an ideal source which possesses no dynamics. The system is shown in Fig. 10. The nonlinear-friction characteristic, denoted by f, is shown in Fig. 11. If the

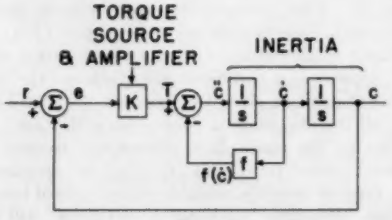


FIG. 10

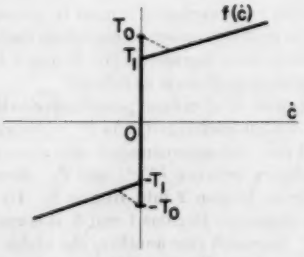


FIG. 11

output element is not moving, torque T_0 is necessary to overcome *static* coulomb friction and start the motion. Once the output element is moving in either direction the coulomb friction drops to its *dynamic* value $T_1 < T_0$ which then remains constant for all $\dot{c} > 0$; to this is added the viscous friction which may be assumed to vary linearly with the velocity $\dot{c}$. Actually, Fig. 11 is a highly idealized picture of a complicated physical situation, about which it is difficult to obtain reliable data. The discontinuity could be lessened by introducing negative slopes as shown in Fig. 11 with dashed lines; this, however, would not affect materially the analysis which follows and is therefore omitted.

The differential equations of the motion in the $(e, \dot{e})$ phase plane may be written down by inspection, assuming that the input is given by Equation [3]

$$\frac{dc}{dt} = \dot{e} \text{ and } \frac{d\dot{e}}{dt} = K(r + \dot{r} - c) - f(\dot{c}) \ldots [7]$$

The simplicity of these equations is treacherous; indeed, Equations [7] only hold when

$$\dot{e} \neq 0 \text{ or } \dot{e} = 0 \text{ and } K|r + \dot{r}t - c| > T_0 \quad \ldots\ldots [8]$$

When $K|r + \dot{r}t - c| < T_0$ no motion can take place because the torque is insufficient to overcome static coulomb friction. The situation becomes easier to visualize when the problem is restated in the $(e, \dot{e})$ phase plane, where the system equations are

$$\left.\begin{aligned} \frac{de}{dt} &= \dot{r} - \dot{c} \text{ and } \frac{d\dot{e}}{dt} = Ke - f(\dot{c}) \text{ when } \dot{e} \neq 0 \text{ or} \\ \dot{e} &= 0 \text{ and } K|e| > T_0 \end{aligned}\right\} \quad \ldots [9a]$$

$$\left.\begin{aligned} \frac{de}{dt} &= \dot{r} \text{ and } \frac{d\dot{e}}{dt} = 0 \text{ when } \dot{e} = 0 \\ &\text{and } K|e| < T_0 \end{aligned}\right\} \quad \ldots\ldots [9b]$$

The critical points of Equations [9a] are at

$$\dot{e} = \dot{r} \text{ and } \frac{1}{K} f(\dot{r}) = e \quad \ldots\ldots\ldots\ldots [10]$$

Eliminating $\dot{r}$ from Equation [10], it is found that

$$\dot{c} = \left(\frac{1}{K} f\right)^{-1}(e)$$

at the critical points; i.e., the critical points lie on the inverse friction curve in the $(e, \dot{c})$ phase plane. Equations [9b] have no critical point.

The critical points given by Equation [10] depend on the parameter $\dot{r}$. There are two cases to be considered.

(i) $\dot{r} = 0$. In the upper half of the phase plane, $\dot{c} > 0$, the trajectories obey a linear differential equation and belong to a virtual critical point located at $e = T_1/K$. By symmetry, the trajectories in the lower-half plane $\dot{c} < 0$ belong to a virtual critical point located at $e = -T_1/K$. If the gain K is sufficiently large, the critical points will be stable foci; if K is sufficiently small, they will be stable nodes. If the critical points are foci, the trajectories spiral around the origin until they intersect the line segment $\dot{c} = 0$ and $|e| \leq T_0/K$, at which instant the motion stops, because the rate of change of both e and $\dot{c}$ is then zero, according to Equation [9b]. In a sense, the entire line segment is a locus of critical points in this case. Physically, the situation is described by the phrase, "The system sticks," i.e., the torque generated by the system is insufficient to overcome static coulomb friction after a certain time. A sketch of a typical trajectory is shown in Fig. 12. Only accidentally can a trajectory reach one of the critical points; hence they are to be regarded virtual in the sense of Section 3.1. Substantially the same result holds when the critical points are stable nodes.

(ii) $\dot{r} \neq 0$. This is a much more interesting case. Here $de/dt = \dot{r} \neq 0$ when the trajectory is on the line segment corresponding to static coulomb friction, so that the trajectory cannot stay there but must leave it at $e = \pm T_0/K$. One of the critical points is always actual, the other virtual. If the critical points are stable nodes, the trajectory tends monotonically to N_1^+ once it has left the static-friction segment and the system is stable for all $\dot{r}$. But if the critical points are foci and $\dot{r}$ is sufficiently *small*, the situation shown in Fig. 13 may arise. In other words, the trajectory leaving the e-axis at $e = T_0/K$ and spiraling around F_1^+ may reintersect the e-axis at $|e| < T_0/K$ and a limit cycle results. Physically speaking, the system alternately sticks and moves and the resulting oscillation is decidedly nonsinusoidal in appearance. The limit cycle may be destroyed by increasing the magnitude of $\dot{r}$, so that the trajectory leaving the e-axis at T_0/K and spiraling around the critical point cannot reintersect the e-axis. This is shown by the dashed trajectory in Fig. 13 which arises when the magnitude of the ramp is $\dot{r}'$. Alternately, for any fixed $\dot{r}$, the limit cycle also can be destroyed by letting the ratio T_0/T_1 approach unity. Thus the instability in such a system (i.e., the limit cycle) results from the interaction of several different elements in the system: (*a*) The discontinuous friction characteristic; in particular, the fact that static coulomb friction exceeds dynamic coulomb friction; (*b*) the input must be of particular type, a small ramp in this

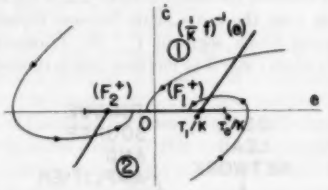


FIG. 12

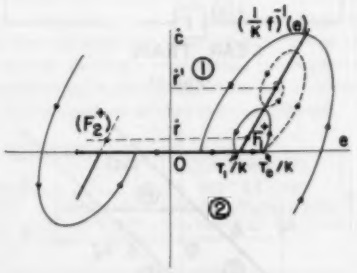


FIG. 13

case; (*c*) the gain must be such that the actual critical point is a focus.

Further details of the problem are given by Lauer (14), who discusses it from a somewhat different point of view.

The oscillations encountered here are frequently called "relaxation oscillations," by which is meant that ares of the limit cycle are contained in regions of the phase space where the motion is subject to different physical laws. It is significant that the instability can only be brought about by a special type of input, and is absent when the system is subject to other inputs. Extrapolating this observation suggests that many control systems with sufficiently strong nonlinearities may exhibit instabilities when subjected to a small, special class of inputs, which perhaps, are present only with a vanishingly small probability in the class of all inputs for which the system is designed. For instance, certain control systems behave in an unsatisfactory fashion when the input is periodic (or a random signal with a very strong periodic component), while they perform quite well when the inputs are of a less special kind. Unfortunately, no serious study of such questions exists at this time.

5 Instability Caused by Multivalued Nonlinearities

Interesting nonlinear effects arise when the trajectories starting at any point in the phase plane are not necessarily unique, but may depend on the "state" of the nonlinearity. Possibly the simplest example of such effects is afforded by backlash (dead-zone in gear trains).

Example 5. Consider another simple positional servomecha-

nism, shown in Fig. 14, similar to that discussed in connection with Example 4. The difference is that now the output element is purely inertial (with negligible friction); an ideal lead network is used to stabilize the system. The nonlinearity arises because, to obtain the required feedback, a gear train must be used which is (as all conventional gear trains are) subject to backlash. The backlash characteristic is idealized as shown in Fig. 15. It is subject to the following interpretation: If $\dot{c} > 0$, i.e., the gears move in the positive direction, segment A of the characteristic applies. If now $\dot{c}$ decreases and goes through zero, then $f(c)$ will stop at the value reached when the velocity was just zero (segment B), until c has traveled the small distance 2Δ in the negative direction, at which time the gears again become enmeshed and $f(c)$ starts decreasing along segment C. By symmetry, the same process occurs when $\dot{c}$ starts to increase, going through zero. The

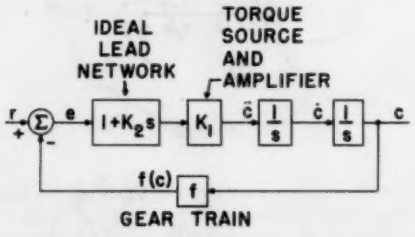


FIG. 14

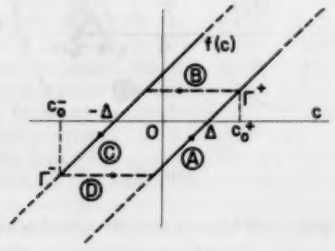


FIG. 15

slopes of segments A and C may be taken as unity. The differential equations of the system in the $(c, \dot{c})$ phase plane are, assuming $r(t) \equiv 0$

$$\frac{dc}{dt} = \dot{c} \text{ and } \frac{d\dot{c}}{dt} = -K_1(K_2\dot{c} + c - \Delta) \quad [11a]$$

provided $\dot{c} > 0$ and gears are enmeshed (segment A)

$$\frac{dc}{dt} = \dot{c} \text{ and } \frac{d\dot{c}}{dt} = -K_1\Gamma^+ \quad [11b]$$

provided $\dot{c} < 0$ and $c_0^+ > c > c_0^+ - 2\Delta$, and gears are not enmeshed (segment B)

$$\frac{dc}{dt} = \dot{c} \text{ and } \frac{d\dot{c}}{dt} = -K_1(K_2\dot{c} + c + \Delta) \quad [11c]$$

provided $\dot{c} < 0$ and gears are enmeshed (segment C)

$$\frac{dc}{dt} = \dot{c} \text{ and } \frac{d\dot{c}}{dt} = -K_1\Gamma^- \quad [11d]$$

provided $\dot{c} > 0$ and $c_0^- < c < c_0^- + 2\Delta$, and gears are not enmeshed (segment D).

Terms c_0^+, c_0^-, Γ^+, Γ^-, Δ are defined in Fig. 15.

The regions of applicability of Equations [11a] and [11d] overlap; the same holds for Equations [11b] and [11c]. To simplify the discussion, assume arbitrarily that the gears are initially enmeshed and will continue to be for a small subsequent period of time. This amounts to specifying the initial state of the nonlinearity.

(i) $r(t) \equiv 0$. There are two linear families of trajectories corresponding to the segments A and C in Fig. 15. For any fixed K_2, the critical points will first be stable foci and then stable nodes as K_1 increases. Let the critical points be nodes; they are located at $\dot{c} = 0$ and $c = \pm\Delta$.

What happens to a trajectory tending to one of the nodes, say N_A^+), when it intersects the c-axis (see Fig. 16)? Assume the intersection occurs at $c > \Delta$. The subsequent trajectory will obey [11b], until c has decreased by the amount 2Δ, at which time [11c] becomes applicable (segment C in Fig. 15). The family of trajectories obeying [11b] is not linear because the forcing function Γ^+ depends on the point where the trajectory intersects the c-axis. The locus of all points on trajectories of the family [11b] whose c-co-ordinate is at a distance 2Δ from the

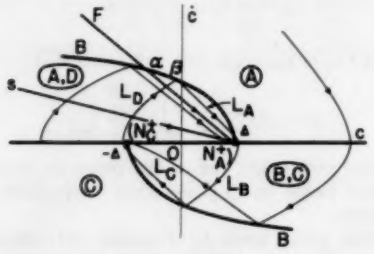


FIG. 16

point where the trajectory crosses the c-axis, is indicated by the heavy curve B in the phase plane. This curve may be shown to consist of parabolic segments.

It is now a simple matter to sketch representative trajectories. If the trajectory starts in the upper [lower] half plane, it will tend to N_A^+) [$(N_C^+]$ until it arrives at the critical point or until it intersects the c-axis from above [below]. The trajectory is then governed by [11b] (or [11d]) until it intersects the boundary curve B, after which it will tend to $(N_C^+ [N_A^+)]$ and so on.

Since the regions of validity of [11a to d] overlap, more than one trajectory may pass through the same point in the phase plane. This complication may be removed by using *two* phase planes; (a) a plane corresponding to the state, "gears not enmeshed," in which the trajectories obey [11b] and [11d]; (b) a plane corresponding to the state, "gears enmeshed," in which the trajectories obey [11a] and [11c]. The two planes are to be joined along the c-axis and along the curve B, but otherwise have no points in common. Hence Fig. 16 may be recognized as the projection of the trajectories of plane (a) on plane (b), subject to the assumption that all trajectories start initially in plane (b). The scheme just described is really a special case of a three-dimensional phase space, where the third dimension has been quantized so that only two values, corresponding to the states of the nonlinearity, are of interest. Mathematically, the situation is entirely analogous to the use of Riemann surfaces for representing multivalued functions.

Finally, from the geometry of the trajectories belonging to the nodes, it is clear that the trajectories cannot intersect the c-axis at $|c| < \Delta$. What is the behavior of trajectories near the critical points? There are two straight-line trajectories tending to each node, such as F and S tending to N_A^+). These are so-called eigenvectors of the linear system (4) associated with the node.

All trajectories which ultimately arrive at N_A^+) must approach it in the sector bounded by the c-axis and F and containing S; all such trajectories tend asymptotically to S. On the other hand, if trajectories in the upper half plane originate outside of this sector, ultimately they will intersect the c-axis, move into the lower half plane, etc. Hence N_A^+) appears as an actual critical point for all trajectories contained within the former sector, and it appears as a virtual critical point for trajectories not contained in that sector. Therefore the nodes may be called semi-actual or semi-virtual; this is the reason for notation N_A^+), (N_C^+.

Notice also that it is possible to have a stable limit cycle $L_AL_BL_CL_D$, as shown in Fig. 16. All trajectories which intersect the c-axis in the interval $\Delta < |c| < c_L$, where c_L denotes the intersection of the limit cycle with the positive c-axis, will tend to the limit cycle. Further, if a trajectory is at or near a critical point, a small perturbation is sufficient to carry it to the limit cycle. Hence, practically speaking, all trajectories will ultimately lead to the limit cycle.

From Fig. 16 it is seen that a necessary condition for the existence of a limit cycle is that it intersect the curve B at a point β which is closer to the origin than the intersection α of the line F with B. By a much more complicated analysis it is possible to show that this condition is also sufficient for the existence of the limit cycle (17), which is stable. Since from [11a] the equation of the line F is (4)

$$\dot{c} = s_2(c - \Delta) \quad \ldots\ldots\ldots\ldots\ldots\ldots [12]$$

where s_2 is the smaller (in algebraic value) characteristic root of [11a] and since $s_2 \to -\infty$ as $K_1 \to \infty$, it is not possible to avoid the situation shown in Fig. 16 by any choice of K_1, although the amplitude of the limit cycle can be somewhat decreased by increasing K_1. To destroy the limit cycle altogether, it is necessary to prevent F from intersecting B, which can be done only by modifying B. One possible solution lies in providing damping for the output element of the servo (17).

Analysis of the case where the critical points of [11a] and [11c] are foci (17, 18) is similar. Then both critical points are virtual; a stable limit cycle always exists.

Owing to the double-valued nonlinearity, the phase-plane structure shown in Fig. 16 is quite special. For instance, it violates the well-known requirement, valid when the nonlinearities are single-valued as in Section 3, that a stable limit cycle must enclose an unstable limit cycle or an unstable critical point—the stable limit cycle in Fig. 16 encloses two semi-actual critical points. This irregularity of the phase-plane structure may explain a recent interesting result of Nichols (15), who found that if a system similar to that considered here is analyzed by means of Kochenburger's method (11), a stable and an unstable limit cycle are predicted under certain conditions, whereas in reality only a stable limit cycle exists. This failing of the describing-function method is probably due to the fact that it calls for "linearizing" the nonlinearity of Fig. 15 in such a fashion which does not properly take into account the discontinuities. A safer procedure would consist in considering additional dynamic effects, such as inertia and resilience of the gearing, in which case the nonlinearity becomes single-valued (16), but then it is necessary to deal with a higher-dimensional phase space and the situation will be appreciably more complicated.

(ii) $r(t) \neq 0$. If the input is allowed to be a step, but not a ramp, the preceding discussion remains valid with only trivial changes in notation. If a ramp input is admitted, however, there is an essential difference. In that case, the critical points will be located at

$$\dot{c} = \dot{r} \qquad r - c = \pm\Delta \ldots\ldots\ldots\ldots\ldots [13]$$

This means that one of the critical points will be actual, the other virtual. As a result, the limit cycle will be destroyed. Physically, this is exactly what one should expect; if the gears are always driven in one direction, the effects of backlash disappear.

6 Instability Arising From the Geometry of Phase Space

So far, the phase space of systems studied, excepting Example 2, was a plane; i.e., two-dimensional and infinite in both dimensions. The next example is a system defined over a *cylindrical phase space*. The phase space is still two-dimensional but it is infinite only in one of the dimensions.

Example 6. Consider a simple positional servomechanism shown in Fig. 17. The distinguishing feature of this system is that the output is detected by means of a continuous-rotation potentiometer. In other words, after rotation by the angle 2π, the wiper arm of the potentiometer returns to its original position.

To describe a nonlinearity of this type analytically and to set up the corresponding differential equations would be somewhat awkward. The following simple device, commonly used in topology (19), eliminates the difficulty and permits simple visualization of the problem. When $|e| < \pi$, the system obeys a simple linear differential equation in the (e, $\dot{e}$) phase plane. Now the fact that rotation by $c = 2\pi$ does not change the physical situation can be expressed by identifying points on the line $c = -\pi$ with points on the line $c = \pi$ having the same $\dot{c}$-co-ordinate. Such an identification is merely another way of stating that the phase space is the (surface of the) infinite cylinder. The reader

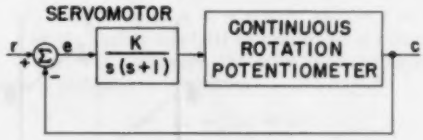


Fig. 17

may easily verify this state of affairs by cutting out an elongated rectangular strip and pasting together, i.e., identifying the long sides.

Notice that the statement, "The phase space is cylindrical," uniquely characterizes the nonlinearity involved, namely, the continuous-rotation potentiometer. A system defined over a cylindrical phase space is necessarily nonlinear.

To analyze the system more closely, assume that the input is a step of magnitude $|r|$. There is only one critical point at

$$c = r \text{ and } \dot{c} = 0 \quad \ldots\ldots\ldots\ldots\ldots\ldots [14]$$

Depending on the loop gain K, the critical point may be a stable node N^+ or a stable focus F^+. When the cylinder is cut along $c = \pi$ and its surface is spread out on the plane, all trajectories are governed by a linear differential equation. There are several interesting cases:

(i) $|r| > \pi$. In this case there exists no critical point, for the point $c = r$, $\dot{c} = 0$ cannot be found on the (surface of the) cylinder. Of course, this is merely the result of stupid design; the servo cannot be expected to work with such large inputs. The problem of how the servo will behave is still interesting, however; it is also easy to explore. Let the critical point be a stable node; since it is located "outside" the phase space, it may be called a virtual critical point and denoted by (N^+). The corresponding trajectories are shown in Fig. 18. A representative trajectory is denoted by $ABB'CC'DD'$. The arc $D'D$ is a limit cycle. Since in the phase space all trajectories ultimately converge to the eigenvector (4) ED, it is clear that there is only one limit cycle, $D'D$,

and that all trajectories tend to it. Physically, the limit cycle corresponds to the motor driving the potentiometer always in the same direction, but with a periodically repeated nonconstant speed. The case when the virtual critical point is (F^+) is essentially the same.

The limit cycle may be visualized by imagining a rubber band stretched around the cylinder in such a fashion that if the cylinder is cut along the rubber band, it falls into two half-cylinders. The limit cycle $D'D$ is called a *limit cycle of the second kind* (2–3) to distinguish it from the limit cycle of the first kind, which arose in connection with Examples 1 through 5. If the cylinder were cut along a limit cycle of the first kind, the result would be a circular patch and a cylinder with a hole in its surface.

(ii) $|r| < \pi$. If the critical point is N^+, a glance at the foregoing analysis shows that there can exist no limit cycle and that all trajectories tend to the critical point. Suppose, however, that the gain K is large and therefore the critical point is F^+. point. A new critical point has arisen when the differential equation is defined on a new type of surface! This can be explained very simply with the aid of topology. According to a fundamental theorem of Poincaré, the sum of indexes[3] of critical points of a surface must be equal to the Euler characteristic of the surface χ. The number $\chi = 1$ for the plane if the system is incrementally stable at every point sufficiently far from the origin; under the same assumption (which is always true in Example 6), χ (cylinder) = 0. Hence either (*a*) there are no critical points on the cylinder or (*b*) the number of saddle points is equal to the number of other types of critical points. Case (i) of this example illustrates (*a*) and Case (ii) illustrates (*b*). A similar situation exists if the differential equation is defined on the torus (i.e., when both phase-plane variables have the periodic property), since χ (torus) = 0 *always* (19).

A trivial example of a nonlinear system defined over a cylindrical phase space is the ordinary pendulum. The most common

TABLE 1 SUMMARY OF EXAMPLES

Example	Principal cause of instability	Conditions for stability — Sufficient	Conditions for stability — Necessary	Nature of instability	Effect of input
1–3	Root locus crosses over into half plane Re(s) > 0	No roots with Re(s) > 0 for any incremental gain	Not known. Likely to be extremely complicated	Stable and unstable limit cycles; paths to infinity	May alter stability of critical point or limit cycles
4	System equations are defined discontinuously	System overdamped when operating linearly (low gain)	Overdamped: necessary only when $\dot{r} \not\equiv 0$ and small; $T_0 > T_1$	Relaxation oscillation (nonanalytic limit cycle)	Sufficiently small ramp input creates limit cycle
5	Multivalued nonlinearity	Viscous damping of output element		Stable limit cycle	Ramp input destroys limit cycle
6	Geometry of phase space	$\lvert r\rvert < \pi$ and overdamped (low gain)	$\lvert r\rvert < \pi$ if $K \leq 1/4$ $\lvert r\rvert < \pi \tanh \pi/4K\sqrt{(1 - 1/4K)}$ if $K > 1/4$	Stable limit cycle of second kind	Sufficiently large step input creates limit cycle

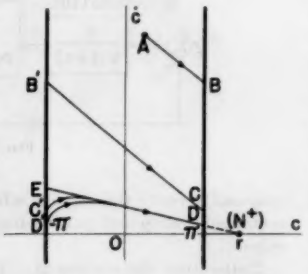


FIG. 18

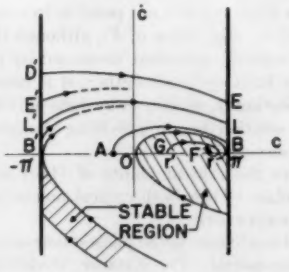


FIG. 19

Furthermore, let the critical point be near one of the boundaries of the phase space, as in Fig. 19. If the trajectory originates at G near the critical point, then the transient is stable. This will happen, for instance, when there is a slight increase in input from r' to r. But if the trajectory originates much farther away, say at A, then it cannot spiral into the critical point without intersecting the line $c = \pi$ at B. Therefore the trajectory continues from B', etc., asymptotically approaching the limit cycle $L'L$. All trajectories starting above the limit cycle, such as $D'EE' \ldots$, approach the limit cycle asymptotically from the other side. The limit cycle is again of the second type.

The region in which the trajectories converge to F^+ is shown by the crosshatching in Fig. 19. The boundary of the region is a trajectory. On the cylinder, the cross-hatched region appears as a helical band. As the critical point approaches either π or $-\pi$, the crosshatched region will shrink, becoming a point when $|r| = \pi$. To assure stability for all $|r| < \pi$, it is desirable that the critical point be N^+. But if the range of the magnitude of r is smaller than π, then somewhat less than critical damping can be tolerated (see Table 1).

It is interesting to observe from Fig. 19 that the point $(\pm\pi, 0)$ on the surface of the cylinder is a critical point; namely, a saddle

example of engineering interest is the synchronous motor. No doubt there exist many other systems, especially among those connected with navigation and guidance problems, where one (or more) dimension in the phase space is not infinite but "cylindrical." When that is the case, effects similar to those discussed may be anticipated.

7 Instability Caused by Sampling

A new type of discontinuity appears when a control system does not operate on continuous signals but receives information intermittently, in "sampled" form. If all other elements of the system are linear, sampling will not render it nonlinear, in fact, the linear theory of sampled-data systems is now well established (20). If there is some nonlinear element in the system, for instance, a relay or a saturating amplifier, then the interaction between the nonlinearity and the discontinuity due to sampling may lead to very complex phenomena which are as yet only incompletely understood. In view of the difficulty of the problem, no discussion is attempted here and the reader is referred to a forthcoming paper of the author (21).

[3] The index of nodes and foci is +1; the index of a saddle point is −1 (see references 2, 3).

Conclusion

The salient features of the examples discussed in the foregoing are collected in Table 1, which requires no further comment.

There is little doubt that such a table is potentially incomplete at the present time, and that it will be superseded in the future by a richer as well as better organized classification of principal nonlinear phenomena encountered in control systems. The author would be greatly appreciative of any communication as well as criticism of the table, particularly welcoming practical examples which appear to violate or to confirm the contents of the table.

Acknowledgments

This research was supported in part by the U. S. Air Force under Contract AF 30(635)-2815, Task No. III, 45360, Project No. 4506. The paper has gained much from conversations of the author with Dr. L. A. Zadeh and Dr. K. Klotter.

References

1 "Phase-Plane Analysis of Automatic Control Systems With Nonlinear Gain Elements," by R. E. Kalman, Trans. AIEE, vol. 73, part II, 1954, pp. 383–390.

2 "Theory of Oscillations," by A. A. Andronov and C. E. Chaikin, translated by S. Lefschetz, Princeton University Press, Princeton, N. J., 1949.

3 "Introduction to Nonlinear Mechanics," by N. Minorsky, J. W. Edwards, Ann Arbor, Mich., 1947.

4 "Analysis and Design Considerations of Second and Higher Order Saturating Servomechanisms," by R. E. Kalman, Trans. AIEE, vol. 74, part II, 1955, pp. 294–310.

5 "Engineering Cybernetics," by H. S. Tsien, McGraw-Hill Book Company, Inc., New York, N. Y., 1954.

6 "Synthesis of Feedback Control Systems by Phase-Angle Loci," by Y. Chu, Trans. AIEE, vol. 71, part II, 1952, pp. 330–339.

7 "The Study of Transients in Linear Feedback Control Systems by Conformal Mapping and the Root Locus Method," by V. C. M. Yeh, Trans. ASME, vol. 76, 1954, pp. 349–361.

8 "Phase-Plane Analysis—A General Method of Solution of Two-Position Process Control," by D. P. Eckman, Trans. ASME, vol. 76, 1954, pp. 109–116.

9 "A General Equation for Relaxation Oscillations," by N. Levinson and O. K. Smith, *Duke Math. Journal*, vol. 9, 1942, pp. 382–403.

10 "Limit Cycles and Rotated Vector Fields," by G. F. D. Duff, *Annals of Mathematics*, vol. 57, 1953, pp. 15–31.

11 "A Frequency-Response Method for Analyzing and Synthesizing Contactor Servomechanisms," by R. J. Kochenburger, Trans. AIEE, vol. 69, 1950, pp. 270–284.

12 "The Use of Nonlinear Feedback to Improve the Transient Response of a Servomechanism," by J. B. Lewis, Trans. AIEE, vol. 71, part II, 1952, pp. 449–453.

13 "A Differential Analyser Study of Certain Nonlinearly Damped Servomechanisms," by R. R. Caldwell and V. C. Rideout, Trans. AIEE, vol. 72, part II, 1953, pp. 165–170.

14 "Operating Modes of a Servomechanism with Nonlinear Friction," by H. Lauer, *Journal of The Franklin Institute*, vol. 255, 1953, pp. 497–511.

15 "Backlash in a Velocity-Lag Servomechanism," by N. B. Nichols, Trans. AIEE, vol. 72, part II, 1953, pp. 462–467.

16 "Stability Characteristics of Closed-Loop Systems With Dead Band," by C. H. Thomas, Trans. ASME, vol. 76, 1954, pp. 1365–1382.

17 "Instrument Inaccuracies in Feedback Control Systems with Particular Reference to Backlash," by H. T. Marcy, M. Yachter, J. Zauderer, Trans. AIEE, vol. 68, 1949, pp. 778–788.

18 "An Introduction to the Analysis of Nonlinear Closed-Cycle Control Systems," by W. E. Scott, "Automatic and Manual Control," edited by A. Tustin, Academic Press, New York, N. Y., 1951.

19 "Combinatorial Topology of Surfaces," by R. C. James, Math. Magazine, vol. 29, 1955, pp. 1–39.

20 "The Analysis of Sampled-Data Systems," by J. R. Ragazzini and L. A. Zadeh, Trans. AIEE, vol. 71, part II, 1952, pp. 225–234.

21 "Nonlinear Aspects of Sampled-Data Control Systems," by R. E. Kalman, presented at Symposium on Nonlinear Circuit Analysis, Polytechnic Institute of Brooklyn, April 26, 1956; to be published in the Proceedings of the Symposium.

Discussion

R. W. Bass.[9] This paper is an exceptionally fine contribution to the subject. It abounds in deep insights and the author displays an acute feeling for the mathematical questions involved. In the writer's opinion, the author's rapid new method of "virtual" and actual critical points provides one of the most valuable techniques now available.

A promising aspect of this method is its applicability in n-dimensions ($n > 2$). But then if a trajectory is trapped between (say) two concentric spherical surfaces with no critical point available, one cannot assert (as in Figs. 5, 7) that the trajectory will tend to a periodic orbit (or limit cycle). In fact, the Poincaré-Bendixson theorem (1)[10] does not hold for $n > 2$. It is true, however, that (2) the trajectory will tend to a limit trajectory which is recurrent in the sense of G. D. Birkhoff (3). A recurrent orbit is nearly "almost periodic in the sense of Bohr." But a recurrent orbit represents unstable performance just as much as does a periodic orbit. Consequently one can avoid the mathematical subtleties (2) involved here, and proceed to use the author's method with success.

The writer's criticism is that not all of the author's statements are established in quite the rigorous style which is currently accepted as mathematical proof. However, the writer has supplied several such proofs and will publish them elsewhere.

In particular, the theorem and corollary of Section 3.1 are near to special cases of an unpublished theorem which the writer had found previously and independently:

Let $x = (x_1, \ldots, x_n)$ be a (column) n-vector, let $|x| \equiv (x_1^2 + \ldots + x_n^2)^{1/2}$, let A be an $n \times n$ matrix, and let I be the unit vector. Let the matrix $\exp(At)$ be defined as in references (4) or (1). Then, if x represents the error (and its derivatives), every control problem can be written

$$\dot{x} = Ax + f(x) + g(t) \ldots\ldots\ldots\ldots [E_1]$$

where, if the reference variable and its derivatives are bounded, $|g(t)| \leq m$ for all $t \geq 0$. We assume that the vector f satisfies $f(0) = 0$.

If $f(x)$ saturates, then we can write $f(x) = \alpha(x)x$, where the matrix $\alpha(x)$ satisfies

$$|\alpha(x_0)x| \leq K|x|$$

for all x and all x_0.

If $\dot{x} = Ax$ is "stable," i.e., if all characteristic roots of A have negative real parts, then

$$|\exp(At)x_0| \leq \gamma\, |x_0| \exp(-\lambda t)$$

for all x_0, all $t \geq 0$, and some $\lambda > 0$, $\gamma > 0$.

The author's criterion is that $\dot{x} = [A + \alpha(x_0)]x$ also be stable for every x_0. A sufficient condition for that is

$$\gamma K < \lambda$$

which we shall assume.

Under these hypotheses, every solution of $[E_1]$ satisfies

$$|x(t)| \leq \gamma[|x(0)| \exp(-\theta t) + m/\theta],\ \theta \equiv \lambda - \gamma K \ldots [I_1]$$

for all $t \geq 0$. Hence any periodic solutions must satisfy

$$|x(t)| \leq \gamma m/\theta \ldots\ldots\ldots\ldots\ldots [I_2]$$

The Inequalities $[I_1]$ — $[I_2]$ recapitulate most of linear and quasilinear servo theory: For step inputs ($m = 0$) the error

[9] Department of Mathematics, Princeton University, Princeton, N. J.

[10] Numbers in parentheses refer to the References at the end of this discussion.

$|x(t)| \to 0$ as $t \to \infty$; while for dynamic inputs the maximum steady-state error depends on the maximum rate of change of the input (measured by m), and the "amount" (θ/γ) of inherent stability.

Corollary: When $m = 0$, a necessary condition for instability is $|f(x_0)| \geq \lambda|x_0|/\gamma$ for some x_0.

We consider next whether the author's piecewise-linear approximations are valid. Kaplan (5) already has proved that frequently one may establish all the qualitative features of the trajectories (the "phase portrait") by examining the field of directions defined by the differential equation at a finite (but "sufficiently large") number of points in the phase space.

Concerning the author's procedure itself, the writer has proved the following theorem:

Let $n = 2$; let the 2-vector function $h(x) = [h_1(x_1, x_2), h_2(x_1, x_2)]$ satisfy a Lipschitz condition. Suppose that

$$\dot{x} = h(x) \quad \ldots\ldots [E_2]$$

is "dissipative," i.e., that there is a circular domain D: $x_1^2 + x_2^2 \leq R^2$ which all trajectories eventually enter. Suppose that $h(0) = 0$ but that $|h(x)| \neq 0$ for $|x| \neq 0$. Then D can be decomposed into the union of certain ("sufficiently small") closed regions $D_i (i = 1, \ldots, N)$ to which corresponds a piecewise-linear, continuous-vector function

$$h_1(x) = A_i x \text{ for } x \text{ in } D_i \ (i = 1, \ldots, N) \quad \ldots\ldots [E_3]$$

which is such that the phase portraits of $[E_2]$ and

$$\dot{x} = h_1(x) \quad \ldots\ldots [E_4]$$

are topologically equivalent (with a single possible exception). The exception: If $[E_2]$ has one or more semistable (i.e., stable-unstable) limit cycles, they may not occur in $[E_4]$. The author's discussion reminds us that stable limit cycles can be observed directly, and unstable limit cycles, indirectly. But (because of "noise") a semistable limit cycle cannot be observed empirically (or graphically) either directly or indirectly. Hence such omissions in the author's approximations $[E_3] - [E_4]$ are of no significance for engineering.

In reply to the author's concluding question, the writer refers to an analysis [(6), summarized in (7), (8)] of the Example 4, in which he neglects coulomb friction, but considers relay dead zone, hysteresis, and time delay. As these three parameters are varied, one can obtain both the author's Fig. 5 and Fig. 7. Since the corresponding ideal system is stable for step inputs, the writer would make a sixth row in Table 1 for this example, listing as the principal cause of instability not "discontinuity" but "system delays." [Incidentally, a new method, based on reference (9), for eliminating such delays is given in (7), (8).] Indeed, delays are a very important source of instability. They could well be listed separately from row three in Table 1, since, e.g., hysteresis does not alter the phase-portrait's independence of time.

REFERENCES

1 "Lectures on Differential Equations," by S. Lefschetz, Princeton University Press, Princeton, N. J., 1946.

2 "Topological Dynamics," by W. H. Gottschalk and G. A. Hedlund, American Mathematical Society Colloquium Publication, vol. 36, 1955.

3 "Dynamical Systems," by G. D. Birkhoff, American Mathematical Society Colloquium Publication, vol. 9, 1927.

4 "Stability Theory of Differential Equations," by R. Bellman, McGraw-Hill Book Company, Inc., New York, N. Y., 1953.

5 "Dynamical Systems With Indeterminacy," by W. Kaplan, *American Journal of Mathematics*, vol. 72, 1950, pp. 573–594.

6 "The Analysis and Synthesis of Relay and Nonlinear Servomechanisms," by R. W. Bass, The Johns Hopkins University Institute for Cooperative Research, 1955.

7 "Improved On-Off Missile Stabilization," by R. W. Bass, American Rocket Society Preprint No. 285–56; submitted to *Jet Propulsion*.

8 "Equivalent Linearization, Nonlinear Circuit Synthesis, and the Stabilization and Optimization of Control Systems," by R. W. Bass, to be published in Proceedings of the Symposium on Nonlinear Circuit Analysis, MRI Symposia Series, vol. 6, October, 1956.

9 "A Generalization of the Functional Relation $Y(t + s) = Y(t)Y(s)$ to Piecewise-Linear Difference-Differential Equations," by R. W. Bass, to appear in the *Quarterly of Applied Mathematics*.

DUNSTAN GRAHAM.[11] It has long been accepted that analyses of nonlinear control systems are restricted to a very limited degree of generality. This paper, in which the author classifies and analyzes nonlinear systems in general terms, represents a fresh approach, the importance of which cannot be overemphasized.

Linear-system design has long profited by the "feel" which the engineer can obtain for the problems which he must attack. This paper is the beginning of a feel for nonlinear problems which will enable the engineer to attack them much more successfully.

K. KLOTTER.[12] The writer wishes to comment briefly on three points of the author's very able presentation. All of the comments pertain either to nomenclature or to definitions:

1 The author has introduced the concept of "virtual critical (or singular) points" as contrasted to "real critical points" and he should be congratulated on this very fortunate designation. The designation seems fortunate because the concept is in immediate and complete parallelism to such well-known and long-established concepts as "virtual images" in geometrical optics, "virtual nodes" in theory of torsional vibrations of engine shafts, and many others.

2 Less fortunate, in the writer's opinion, is another expression coined by the author, the designation of a nonlinear system as "strongly nonlinear" if it is capable of self-sustained oscillations, and (by implication) of "weakly nonlinear" if it is not capable of self-sustained oscillations. The choice of words for describing these two cases seems not so good, because "strong" or "weak" ordinarily carry connotations of quantity rather than quality. It is a distinction of quality, however, which the author intends to make. If one would adopt the author's designations, a system described by van der Pol's differential equation

$$\ddot{q} - \epsilon\dot{q}(1 - \alpha^2 q^2) + \kappa^2 q = 0$$

would have to be called "strongly nonlinear" however small the coefficient ϵ of the nonlinear term may be, whereas a system described by a differential equation like the following

$$\ddot{q} + 2D\kappa\dot{q} + \kappa^2 q(1 + \mu^2 q^2) = 0$$

would have to be called weakly nonlinear however large the coefficient μ^2 of the nonlinear term may be. It seems obvious that such designations will tend to create confusion or at best stand in permanent need of specific and detailed explanations in order to eradicate connotations which come naturally to a reader's or listener's mind.

Instead of the quantitative designations strongly nonlinear or weakly nonlinear qualitative designation like "essentially nonlinear" or "intrinsically nonlinear" or the like and their opposites may be better. However, those expressions also will stand in need of repeated definition and explanation until they may become generally accepted terms. On the other hand, one may ask whether there is indeed a necessity for coining new expressions

[11] Section Head, Automatic Flight Controls, Lear, Inc., Grand Rapids Division, Grand Rapids, Mich.

[12] Professor of Engineering Mechanics, Stanford University, Stanford, Calif. Mem. ASME.

for those cases. Would "capable of sustained oscillations" and "not capable of sustained oscillations" not be sufficient to describe the difference the author wants to emphasize?

3 The last comment pertains to the author's definition of a critical point as unstable if all the real parts of the characteristic roots associated with the critical point are non-negative. This definition compels the author to call a saddle point "neither stable nor unstable."

The author's definition is in strict and direct contradiction to all usages known to the writer, where a system is called unstable if the real part of even one of the characteristic roots is non-negative (see e.g., the Routh-Hurwitz criteria). In this terminology, a saddle point then is to be called unstable.

It seems highly undesirable to introduce definitions which clash with others that have been used for a long time.

Author's Closure

The author is grateful for the lively and constructive criticisms the paper has produced.

Dr. Bass' remark regarding mathematical rigor is worth additional emphasis. Because the behavior of nonlinear systems may present exceptionally subtle aspects, any conclusions in analyzing them should be arrived at as rigorously as possible. Only rigorously obtained results can be used as a sure starting point for further investigations. On the other hand, excessive preoccupation with complete rigor tends to put road blocks in the development of original ideas which, in their crudest form, always stem from intuitive considerations.

In so far as the present paper is concerned, the principal point where additional rigor is needed is the statement, introductory to the theorem in Section 3.1, that ". . . a nonlinearity may be approximated with a sufficiently large number of straight-line segments so that essential features of the root locus are preserved." More precisely, this means that the topological aspects of the solution of a differential equation are preserved whenever the nonlinearity is replaced by a suitably chosen straight-line approximation. The proof of this conjecture is trivial in the case of a first-order differential equation (22).[18] Dr. Bass' theorem stated in the second part of his discussion disposes of most of the interesting cases when the differential equation is of the second order. However, there are some unsettled questions when the differential equation is higher than second order. The conjecture has strong intuitive appeal, but it is important to have a rigorous proof also, particularly to be able to recognize certain degenerate cases when it may not be true.

The examples in the paper after the first one do not make use of this basic conjecture but deal with straight-line-type nonlinearities only. Under these circumstances the statements made in connection with the various examples can be proved rigorously (using mainly geometric facts concerning linear trajectories) when the differential equation is of the second order; in many cases, the discussion had to be greatly abbreviated for lack of space. The details can be easily supplied by the reader after some acquaintance with the method. When the differential equations are of higher than the second order, the procedure is much more difficult and not yet fully worked out, although several examples have been considered in detail by the author in reference (4) of the paper.

As mentioned in the paper, a rigorous proof of the theorem of Section 3.1 is apparently not yet available. It is possible, however, to have somewhat less general theorems concerning monostability. An example of these is the theorem stated by Dr. Bass in connection with the system $[E_1]$. Since it applies only to cases where the sign of the components of the vector function f is immaterial, the practical applicability of the theorem is severely restricted. By certain modifications it is possible largely to remove this limitation (23, 24). At any rate, theorems of this type (which guarantee monostability in a nonlinear system regardless of the order of the differential equation) have much potential engineering usefulness and provide a fertile area for additional work.

The influence of delays, mentioned by Dr. Bass, was not considered since the paper was restricted to differential equations. Practically speaking, delays are introduced in the analysis of control systems because impulse response of linear parts of the system frequently can be more conveniently approximated by a delay term $e^{-\sigma T}$ and a rational transfer function of low degree than by a rational transfer function of high degree. Since physical systems seldom behave corresponding to a mathematically exact delay, the whole matter concerns the semi-empirical problem of replacing a physical system by an idealized mathematical model. Since very little is known about behavior of nonlinear delay-differential equations, the use of time delay for model building purposes in the nonlinear case is a highly treacherous undertaking. This may be illustrated by the celebrated case of the Goodwin business-cycle model.

This model (25) is as follows:

$$\dot{y}(t) + \sigma y(t) = \phi[\dot{y}(t - T)] \ldots\ldots\ldots\ldots [E_6]$$

where $y(t)$ = national income, T = time delay between investment decisions and corresponding outlays, σ = positive constant, and ϕ = nonlinear induced-investment function, similar to the straight-line approximation in Fig. 3 of the paper.

To be able to analyze this system, Goodwin had to expand the term $\dot{y}(t - T)$ in a Taylor series. Moreover, he had to stop at the second term in the Taylor series in order not to get a differential equation of higher than second order. Under these circumstances Goodwin demonstrated the existence of a unique, stable-limit cycle. However, using the approximate method of equivalent linearization, Bothwell (26) has shown that if the delay term is not approximated by the Taylor series there will be an infinite number of limit cycles with monotonically decreasing period. In fact, there will be roughly as many limit cycles as the number of terms retained in the Taylor expansion. Although Bothwell's method was not rigorous, there is little doubt about its validity; the same conclusions have been checked experimentally by means of an analog computer (27). Of course, this destroys the original intent of Goodwin to find a simple explanation for the business cycle using a nonlinear model.

The lesson to be learned is this: Since the empirical or theoretical justification of using a delay term in the foregoing system is no better than, say, using two terms in the Taylor series, it is probably preferable not to use delay to approximate the response of physical systems. If a Taylor series approximation to the delay is used, as in Example 1 of the paper, the instability would fall under the first line in Table 1.

Professor Klotter's second comment refers to terminology used in the earlier (typewritten) version of the paper. As a result of his recommendation made orally during the Princeton conference, the terms "weakly" and "strongly nonlinear" have been replaced in the present version of the paper by "monostable" and "not monostable," respectively. The present wording (suggested by L. A. Zadeh) is motivated by analogous designations used in electrical engineering practice in connection with multivibrators. Specifically, a monostable multivibrator has one stable equilibrium point, a bistable multivibrator has two, while an astable multivibrator has a single, stable, limit cycle.

The terms "not capable of sustained oscillations" or "capable of sustained oscillations" would not be adequate to cover fully the situation because of the possibility of more than one stable equi-

[18] Numbers in parentheses refer to additions to the References of the paper and appear at the end of this closure.

librium point; moreover, in certain cases, the steady-state behavior of a nonlinear dynamic system cannot even be characterized by the phrases "there are one or more stable equilibrium points and one or more stable limit cycles." The so-called "ergodic" case of solutions of a differential equation defined on the torus falls in this category (28). There are also practical cases where the behavior of a sampled-data system (governed by a nonlinear-difference equation) should be described in the language of probability theory (29). Thus the word "monostable" is desirable to describe the situation where the solutions of a nonlinear differential equation are topologically equivalent to solutions of a stable linear differential equation.

Professor Klotter's third comment touches on matters which could not be discussed in the paper because of space limitations. The words "stable" and "unstable" are commonly used for overall characterization of the behavior of linear autonomous systems. Unfortunately, it appears impossible at present to give a really satisfactory definition of stability in the nonlinear case, because of the tremendous range of possibilities which must be considered (30). In the author's view, the use of these concepts in connection with nonlinear systems should be avoided if not altogether excluded because of their inherent ambiguity.

It is sometimes suggestive, however, to use "stable" and "unstable" in a much narrower sense. In particular, the terminology "stable focus" or "unstable focus," and so on, is motivated by the fact that all solutions starting in a small neighborhood of a stable focus converge to it as $t \to +\infty$, while all solutions converge to an unstable focus as $t \to -\infty$. Hence there is a sort of complementarity between stable and unstable focus, corresponding to a change in sign of t. On the other hand, topological aspects of the behavior near a saddle point do not change as t is replaced by $-t$. For this reason, it does not make sense to talk about stable or unstable saddle points. The terminology is intended to accentuate the difference between a foci and nodes on the one hand, and saddle points on the other, which is all important in nonlinear problems. In fact, saddle points are hardly differentiated from unstable foci or nodes in the discussion of linear systems, since they are all "unstable" (in Professor Klotter's terminology). The fact remains that our present-day usage in connection with linear systems is much too oversimplified to accommodate the manifold distinctions which must be recognized in dealing with nonlinear systems. The situation is perhaps not unlike concepts of grammar which are difficult to grasp clearly when studied from the point of view of the English language which has an oversimplified structure, while the same concepts of grammar appear to be much clearer when studied in relation to a more complex language, such as German or Greek.

REFERENCES

22 Appendix A, reference (21).

23 Reference (8) of Dr. Bass' discussion.

24 Oral communication by Dr. Bass.

25 "The Nonlinear Accelerator and the Persistence of Business Cycles," by R. M. Goodwin, *Econometrica*, vol. 19, 1951, pp. 1–17.

26 "The Method of Equivalent Linearization," by F. E. Bothwell, *Econometrica*, vol. 20, 1952, pp. 269–283.

27 "Goodwin's Nonlinear Theory of the Business Cycle: An Electro-Analog Solution," by R. H. Strotz, J. C. McAnulty, and J. B. Naines, Jr., *Econometrica*, vol. 21, 1953, pp. 390–411.

28 "Theory of Ordinary Differential Equations," by E. A. Coddington and N. Levinson, McGraw-Hill Book Company, Inc., New York, N. Y., 1955.

29 Theorem 5, reference (21).

30 "On the Stability of Mechanical Systems," by J. J. Stoker, *Communications on Pure and Applied Mathematics*, vol. 8, 1955, pp. 133–142.

Determination of the Characteristics of Multi-Input and Nonlinear Systems From Normal Operating Records[1]

By T. P. GOODMAN,[2] CAMBRIDGE, MASS.

A method is presented for discovering, from the random variations in normal operating records, the impulse responses, or weighting functions, relating an output of a system to two or more inputs that are mutually correlated. This method makes use of the statistical autocorrelation and cross-correlation functions of the system inputs and outputs. From these correlation functions, the weighting functions are obtained by a process of deconvolution by means of an electronic delay-line synthesizer. The method is described first for linear systems and is then applied to nonlinear systems. It is an extension of a method presented in an earlier paper for discovering the weighting function of a linear system with one input.

Introduction

IN designing, improving, or evaluating the automatic control of a physical system or process, the characteristics of the system must be known in such a form that the output for a given input can be determined. The system characterization to be used in the present study is the impulse response, or weighting function, of the system; that is, the response of the system to a unit impulse input. In a linear system, the step response may be obtained by integrating the impulse response, and the frequency response may be obtained by taking the Fourier transform of the impulse response (1).[3] Heretofore the usual way of obtaining these responses has been to interrupt the operation of the system and subject it to the appropriate artificial disturbances (impulses, step functions, or sinusoids). It has been shown recently (2, 3, 4) however, that the same information can be obtained by statistical correlation of the random variations normally present in the system's inputs and outputs, without subjecting the system to any artificial disturbances.

In an earlier paper (3), a statistical method was described for obtaining the weighting function of a linear system with a single input and a single output. Experimental results (3, 4) indicated that the method had sufficient promise to justify further development in two directions:

1 Additional experience in practical application of the method.

2 Extension of the theory to make it applicable to more complex systems.

The present paper is a contribution in the second direction. It is planned to obtain additional experience in the practical application of the method as rapidly as data from industrial processes become available for study.

In the present paper, the statistical method for obtaining the weighting function of a single-input linear system is briefly reviewed, and the method then is extended to provide a straightforward way of dealing with systems with multiple inputs that are mutually correlated. It is shown that when the method is applied to a nonlinear system, the weighting functions obtained are those of the optimum linear representation (in a mean-square sense) of the system. The application of linear representations in design is briefly discussed. Some methods for obtaining the frequency-response functions corresponding to the weighting functions of a system are suggested.

Review of Method for Single-Input Systems

The dynamics of a linear system with a single input $m(t)$ and a single output $c(t)$, shown in Fig. 1, are characterized completely by the system's impulse response, or weighting function. The

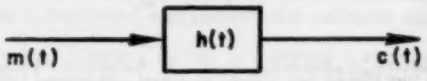


Fig. 1 Single-Input, Single-Output System

weighting function is a continuous function $h(t)$, but for convenience in computation, the area under the curve of $h(t)$ can be approximated by a sequence of thin rectangles, each of width T. Then $h(t)$ for any stable system can be denoted by a finite time sequence

$$h(t) \approx \left[\frac{1}{2}Th(0), Th(T), Th(2T), \ldots, Th(kT)\right]$$

$$\approx [h_0, h_1, h_2, \ldots h_k] \ldots\ldots [1]$$

where the general term h_n is equal to T times the response at time t to a unit impulse at time $(t - nT)$.

As shown in reference (3), the output of the system for any input may be approximated in terms of the weighting function as

$$c(t) = \sum_{n=0}^{k} h_n m(t - nT) \ldots\ldots\ldots\ldots [2]$$

In this paper the assumption is made that the systems dealt with are time invariant; that is, that their weighting functions do not change with time. The method described in this paper cannot be applied usefully to time-variant systems unless the time variations are slow compared with the length of the input and output records required to determine the weighting functions.

---

[1] Based on a portion of a thesis undertaken in partial fulfillment of the requirements for the degree of Doctor of Science in Mechanical Engineering at Massachusetts Institute of Technology, Cambridge, Mass. This investigation was supported in part by a research grant from the Department of Mechanical Engineering, Massachusetts Institute of Technology.

[2] Assistant Professor of Mechanical Engineering, Massachusetts Institute of Technology. Assoc. Mem. ASME.

[3] Numbers in parentheses refer to the Bibliography at the end of the paper.

Contributed by the Instruments and Regulators Division of The American Society of Mechanical Engineers and presented at the ASME-AIEE Conference on Nonlinear Control Systems, Princeton, N. J., March 26–28, 1956.

Note: Statements and opinions advanced in papers are to be understood as individual expressions of their authors and not those of the Society. Manuscript received at ASME Headquarters, February 3, 1956. Paper No. 56—IRD-17.

The process of determining the weighting function of a system from normal operating records of its input and output is facilitated greatly by using the autocorrelation of the input

$$\phi_{mm}(\tau) = \overline{m(t)m(t+\tau)} \quad [3]$$

and the cross-correlation of input with output

$$\phi_{mc}(\tau) = \overline{m(t)c(t+\tau)} \quad [4]$$

where the bar denotes averaging over all available records.

When both sides of Equation [2] are correlated with $m(t)$

$$\overline{m(t)c(t+\tau)} = \sum_{n=0}^{k} h_n \overline{m(t)m(t+\tau-nT)}$$

or

$$\phi_{mc}(\tau) = \sum_{n=0}^{k} h_n \phi_{mm}(\tau - nT) \quad [5]$$

As described in reference (3) the advantages of using the correlation functions, rather than the input-output records themselves, to determine the weighting function of a linear system are:

1 The effects of noise are largely eliminated, even when the system is part of a closed loop.

2 The autocorrelation of a nonperiodic input has a high central peak and approaches a steady-state value for large positive and negative values of τ; hence Equation [5] is easier to solve for the weighting function h_n than is Equation [2].

The process of determining the weighting function h_n from Equation [5] is known as deconvolution. This may be done numerically by solving a matrix equation obtained by writing Equation [5] for $k+1$ values of τ. Deconvolution is facilitated greatly, however, by an electronic device (3) known as the "delay line synthesizer" (DLS). The DLS of Fig. 2 consists of a tapped electronic delay line with 20 equal time delays each of amount T,

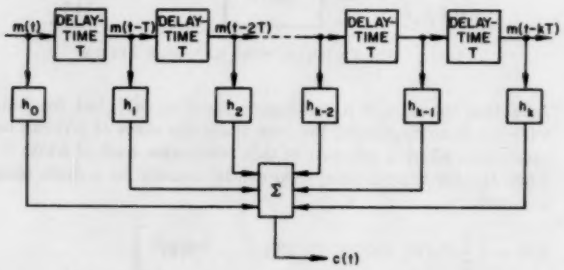


FIG. 2 BLOCK DIAGRAM OF DELAY-LINE SYNTHESIZER (DLS)

coefficient multipliers by which each delayed input is multiplied by the appropriate h_n, and a summing panel. Thus, when $m(t)$ is used as the input voltage to the DLS and the coefficients h_n are set on the multipliers, the output voltage is $c(t)$, as indicated by Equation [2]. Similarly, when $\phi_{mm}(\tau)$ is used as the input, and the same coefficients h_n are used, the output is $\phi_{mc}(\tau)$, as indicated by Equation [5]. Conversely, to solve for the h_n when $\phi_{mm}(\tau)$ and $\phi_{mc}(\tau)$ are known, $\phi_{mm}(\tau)$, as obtained from a function generator, is used as the input, and the coefficient settings are adjusted until the output is matched with $\phi_{mc}(\tau)$ as closely as possible. The weighting function then may be obtained either by reading off the coefficient settings or by experimentally measuring the impulse response of the DLS itself. If the delays obtained by the DLS were perfect, both of these methods would give the same results because the impulse response of the DLS would be a sum of delayed impulses. Since the delays are not perfect, the actual impulse response of the DLS is a sum of rounded pulses, giving a continuous curve with interpolation between the discrete values h_n. For this reason the experimental impulse response of the DLS is a closer approximation than the sequence of coefficient settings h_n. For simplicity, however, the equations of the following discussion are written in terms of the discrete form of the impulse response.

SYSTEMS WITH TWO INPUTS

For a linear system with two inputs and one output, as diagrammed in Fig. 3, there are two weighting functions, which may be denoted by $g(t)$ and $h(t)$, such that

$$z(t) = \sum_{n=0}^{k} [g_n x(t-nT) + h_n y(t-nT)] \quad [6]$$

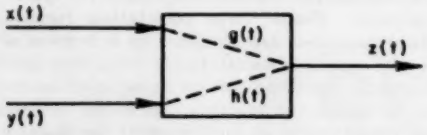


FIG. 3 TWO-INPUT, SINGLE-OUTPUT SYSTEM

To aid in the solution of Equation [6], both sides of the equation may be correlated, first with $x(t)$ and then with $y(t)$

$$\phi_{zx}(\tau) = \sum_{n=0}^{k} [g_n \phi_{xx}(\tau - nT) + h_n \phi_{xy}(\tau - nT)] \quad [7]$$

$$\phi_{yz}(\tau) = \sum_{n=0}^{k} [g_n \phi_{yx}(\tau - nT) + h_n \phi_{yy}(\tau - nT)] \quad [8]$$

A comparison of Equations [7] and [8] with Equation [5] shows that while Equation [5] can be solved numerically by inverting a $(k+1) \times (k+1)$ matrix, the solution of Equations [7] and [8] requires a $2(k+1) \times 2(k+1)$ matrix. However, since the difficulty of solving a matrix equation increases progressively as the number of terms increases, Equations [7] and [8] are much more than twice as difficult to solve as Equation [5].

Equations of the form of Equations [7] and [8] are well known in multiple-correlation analysis and have been used in problems of prediction in economics and in many other fields. Wiener (5) has shown that it is sometimes advantageous to take the Fourier transforms of Equations [7] and [8] to obtain solutions for the weighting functions in the frequency domain before transformation back to the time domain. However, for this investigation, a less cumbersome technique was believed to be desirable.

A simultaneous solution of Equations [7] and [8] can be obtained by use of two DLS's. One DLS is used for the g_n and the other for the h_n. This process requires switching back and forth between two sets of inputs to get two outputs that most nearly approximate ϕ_{zx} and ϕ_{yz}, respectively. If four DLS's are available, and if their coefficients can be coupled to give two pairs of DLS's so that the two DLS's in each pair have the same coefficient settings, then the problem can be solved directly. This procedure is illustrated in Fig. 4. But to avoid the use of so much equipment, it is desirable to have a means for solving for one weighting function independently of the other.

First, it may be noted that if x and y are uncorrelated, that is, if $\phi_{xy}(\tau) = \phi_{yx}(\tau) = 0$ for all τ, then Equations [7] and [8] reduce to equations of the same form as Equation [5], and may be solved separately. This suggests that if x and y are correlated, a trans-

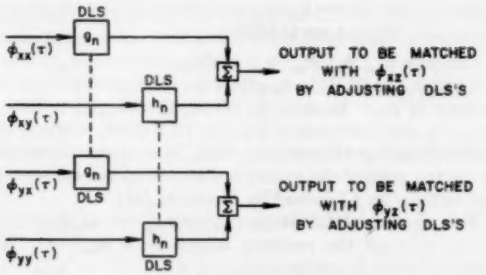


FIG. 4 SIMULTANEOUS DLS SOLUTION FOR WEIGHTING FUNCTIONS OF TWO-INPUT SYSTEM

formation of variables should be sought that isolates the component $w(t)$ of $y(t)$ that is uncorrelated with $x(t)$.

If the relationship between x and y is linear, y may be written in the form.

$$y(t) = w(t) + \sum_{n=-k}^{k} f_n x(t - nT) \quad [9]$$

where the f_n are the time-sequence representation of a linear weighting function relating $x(t)$ and $y(t)$. The summation extends from $-k$ to $+k$ in order to include the effect of the past of $x(t)$ on the future of $y(t)$ (positive n) as well as the effect of the past of $y(t)$ on the future of $x(t)$ (negative n). In the summation, k should be chosen large enough to include all the significant dynamical effects in the physical system; that is, k should be large enough so that

$$\overline{y(t)x(t \pm nT)} \approx 0, \qquad n > k$$

The problem now is to find the f_n that make $w(t)$ uncorrelated with $x(t)$.

When both sides of Equation [9] are correlated with $x(t)$

$$\phi_{xy}(\tau) = \phi_{xw}(\tau) + \sum_{n=-k}^{k} f_n \phi_{xx}(\tau - nT)$$

Thus, by defining the f_n by the relation

$$\phi_{xy}(\tau) = \sum_{n=-k}^{k} f_n \phi_{xx}(\tau - nT) \quad [10]$$

the condition that $\phi_{xw}(\tau) \equiv 0$ is fulfilled.

When the curves ϕ_{xx} and ϕ_{xy} are set up on a function generator, Equation [10] can be solved for the f_n on a DLS. The DLS is then a dynamical model of the weighting function relating $x(t)$ and $y(t)$. "Negative" time delays are obtained merely by shifting the time axis. With the DLS set up in this way, $x(t)$ can be fed into the DLS, and the output, when subtracted from $y(t)$, gives $w(t)$.

This process of finding the component of $y(t)$ that is uncorrelated with $x(t)$ is essentially a process of orthogonalization. The statement that $w(t)$ is uncorrelated with $x(t)$ is equivalent to the statement that $w(t)$ is orthogonal to all the members of the time sequence $x(t \pm nT)$, in the sense that $\overline{w(t)x(t \pm nT)} = 0$ for all n. This use of the word "orthogonal" is an extension of its use in connection with vectors: two vectors are orthogonal when their scalar product is zero.

The weighting function $f(t)$ may be thought of as representing an equivalent causal path relating x and y, Fig. 5. In the actual system, it is reasonable to assume that a nonzero correlation between x and y is an indication of a causal relationship. By using both positive and negative values of n, all linear causal relationships between x and y can be represented mathematically by Fig. 5, even though the physical relationship may be more complex.

When the relationship between x and y is nonlinear, Equation [10] gives the best linear representation (in a mean-square sense) of the nonlinear relationship; this is shown later in the section on

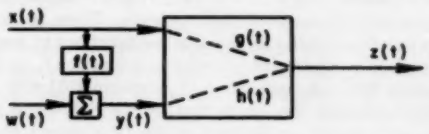


FIG. 5 RELATION BETWEEN INPUTS IN TWO-INPUT SYSTEM

nonlinear systems. For the remainder of this section, this relationship is assumed to be linear.

To simplify the discussion in the remainder of this section, a further assumption is made, that all inputs are statistically stationary; that is, that their statistical properties such as autocorrelations and cross-correlations do not change with time. This assumption permits the time origin to be shifted without affecting the value of the autocorrelation or cross-correlation function. For example

$$\phi_{xx}(\tau + nT) = \overline{x(t)x(t + \tau + nT)} = \overline{x(t - \tau)x(t + nT)}$$

The modifications necessary when dealing with nonstationary inputs are discussed in Appendix 2.

Use of Equation [9] in view of the fact that w is uncorrelated with x, yields Equations [11] through [13]. In these equations, n is merely a dummy variable of summation; any other letter can be used equally well in its place. Hence, in a double summation, it is convenient to distinguish between the two summations by replacing n by m in one of them

$$\phi_{yx}(\tau) = \sum_{n} f_n \phi_{xx}(\tau + nT) \quad [11]$$

$$\phi_{yy}(\tau) = \phi_{ww}(\tau) + \sum_{m}\sum_{n} f_m f_n \phi_{xx}(\tau + mT - nT)$$

$$= \phi_{ww}(\tau) + \sum_{m} f_m \phi_{xy}(\tau + mT) \quad [12]$$

$$\phi_{yz}(\tau) = \phi_{wz}(\tau) + \sum_{n} f_n \phi_{xz}(\tau + nT) \quad [13]$$

When Equations [11] and [12] are substituted in Equation [8]

$$\phi_{yz}(\tau) = \sum_{n} g_n \left[\sum_{m} f_m \phi_{xx}(\tau + mT - nT)\right] + \sum_{n} h_n \phi_{wz}(\tau - nT) + \sum_{n} h_n \left[\sum_{m} f_m \phi_{xy}(\tau + mT - nT)\right] \quad [14]$$

When Equation [7] is substituted in Equation [13]

$$\phi_{yz}(\tau) = \phi_{wz}(\tau) + \sum_{m} f_m \sum_{n} [g_n \phi_{xx}(\tau + mT - nT) + h_n \phi_{xy}(\tau + mT - nT)] \quad [15]$$

Comparison of Equation [14] with Equation [15] shows that

$$\boxed{\phi_{wz}(\tau) = \sum_{n} h_n \phi_{ww}(\tau - nT)} \quad [16]$$

Consequently, Equation [16] is an equation that can be solved

directly for the weighting function h_n. This equation can be solved conveniently on the DLS. By interchanging the roles of x and y in the derivation, a similar equation can be found that can be solved directly for g_n. These weighting functions then can be checked by use of Equations [7] and [8].

To use Equation [16], $\phi_{ww}(\tau)$ and $\phi_{wz}(\tau)$ first must be obtained from Equations [12] and [13], respectively. These latter equations can be solved readily for the required functions by use of the DLS when it is set up to give the weighting function f_n as defined by Equation [10]. A comparison of Equation [10] with Equation [12] shows that the summation on the right-hand side of Equation [12] would be obtained as the DLS output by using $\phi_{xy}(\tau)$ as the input if there were a minus sign instead of a plus sign inside the parentheses. Since the sign is plus, $\phi_{xy}(-\tau)$ is used as the input to the DLS, and the output is then $\phi_{yy}(-\tau) - \phi_{ww}(-\tau)$. The resulting curve may be traced and then subtracted graphically from $\phi_{yy}(-\tau)$ to give $\phi_{ww}(-\tau)$, or the subtraction may be performed electronically, by use of analog-computer elements. In the same way, $\phi_{wz}(-\tau)$ can be found by using Equation [13].

To summarize, as shown in Figs. 6(*a–e*), the steps in finding the weighting function of one input in a system with two inputs x and y and a single output z are as follows:

1 Compute ϕ_{xx}, ϕ_{xy}, ϕ_{yy}, ϕ_{xz}, and ϕ_{yz}.

2 Using ϕ_{xx} as DLS input, adjust the DLS coefficients so that the output is ϕ_{xy}. In order to match this output it may be necessary to shift its τ-axis so that the DLS gives, in effect, time advances as well as time delays. The DLS, when allowance is made for the shift of the τ-axis, is now set up to represent the weighting function f_n, defined by Equation [10].

3 With the same DLS settings, apply $\phi_{xy}(-\tau)$ as input to the DLS and subtract the resulting output from $\phi_{yy}(-\tau)$. This gives $\phi_{ww}(-\tau)$(see Equation [12]).

4 With the same DLS settings, apply $\phi_{xx}(-\tau)$ as input to the DLS, and subtract the resulting output from $\phi_{yx}(-\tau)$. This gives $\phi_{wx}(-\tau)$(see Equation [13]).

5 Using $\phi_{ww}(\tau)$ as DLS input, adjust the DLS coefficients so that the output is $\phi_{wz}(\tau)$. The DLS is now set up to represent the weighting function h_n, and Equation [16] has been solved.

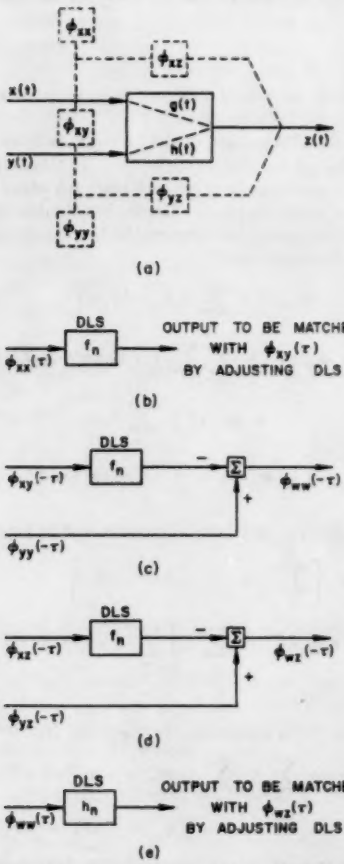


FIG. 6 SUMMARY OF METHOD FOR DETERMINATION OF ONE WEIGHTING FUNCTION OF TWO-INPUT SYSTEM

(*a*) Determination of ϕ_{xx}, ϕ_{xy}, ϕ_{yy}, ϕ_{xz}, and ϕ_{yz}
(*b*) Determination of f_n
(*c*) Determination of ϕ_{ww}
(*d*) Determination of ϕ_{wz}
(*e*) Determination of h_n

SYSTEMS WITH MORE THAN TWO INPUTS

Essentially, what was done in the foregoing section was to "isolate" the weighting function h_n by "isolating" the component of $y(t)$ that is uncorrelated with $x(t)$. This suggests that if there are more than two inputs, the component of one input that is uncorrelated with all the others should be isolated.

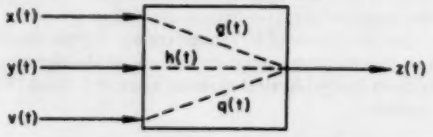


FIG. 7 THREE-INPUT, SINGLE-OUTPUT SYSTEM

To illustrate the procedure when there are more than two inputs, a three-input system is shown in Fig. 7. Now

$$z(t) = \sum_{n=0}^{k} [g_n x(t - nT) + h_n y(t - nT) + q_n v(t - nT)] \quad [17]$$

When both sides of Equation [17] are correlated with x, y, and v, respectively

$$\phi_{xz}(\tau) = \sum [g_n \phi_{xx}(\tau - nT) + h_n \phi_{xy}(\tau - nT) + q_n \phi_{xv}(\tau - nT)] \quad [18]$$

$$\phi_{yz}(\tau) = \sum [g_n \phi_{yx}(\tau - nT) + h_n \phi_{yy}(\tau - nT) + q_n \phi_{yv}(\tau - nT)] \quad [19]$$

$$\phi_{vz}(\tau) = \sum [g_n \phi_{vx}(\tau - nT) + h_n \phi_{vy}(\tau - nT) + q_n \phi_{vv}(\tau - nT)] \quad [20]$$

The component of v which is orthogonal to x and y may be denoted by u. To isolate this component, v may be written in the form

$$v(t) = u(t) + \sum [b_n x(t - nT) + c_n w(t - nT)] \quad [21]$$

with the result that

$$\phi_{ux}(\tau) = \sum b_n \phi_{xx}(\tau - nT) \quad [22]$$

and

$$\phi_{uy}(\tau) = \sum c_n \phi_{ww}(\tau - nT) \quad [23]$$

Equations [22] and [23] can be solved on the DLS in the same manner as Equation [10], and give the weighting functions b_n and c_n when $\phi_{xx}(\tau)$ is obtained from

$$\phi_{xx}(\tau) = \phi_{xx}(\tau) + \sum f_n \phi_{xx}(\tau + nT) \quad [24]$$

in the same way that $\phi_{xx}(\tau)$ is obtained from Equation [12].

Other correlation functions that are needed are

$$\phi_{xx}(\tau) = \phi_{xx}(\tau) + \sum [b_n \phi_{xx}(\tau + nT) + c_n \phi_{xx}(\tau + nT)] \quad [25]$$

$$\phi_{xx}(\tau) = \phi_{xx}(\tau) + \sum [b_n \phi_{xx}(\tau + nT) + c_n \phi_{xx}(\tau + nT)] \quad [26]$$

From Equations [18] through [26] it follows that

$$\boxed{\phi_{xx}(\tau) = \sum q_n \phi_{xx}(\tau - nT)} \quad [27]$$

Thus a means is achieved for solving for the weighting function q_n directly on the DLS when $\phi_{xx}(\tau)$ and $\phi_{xx}(\tau)$ have been found from Equations [25] and [26], respectively. These equations can be solved by use of the DLS in the same way as for Equations [12] and [13].

This procedure may be extended to any number of input variables; for example, if there are four inputs, the first step is to find the weighting functions relating the fourth input to x, w, and u. As the number of variables is increased, however, the amount of work involved increases out of all proportion to the increase in the number of variables. Also, if the correlations are obtained experimentally, they must have high initial accuracy in order to survive a number of subtractions with meaningful results.

Systems With More Than One Output

A system with more than one output introduces no further complications because each output can be treated independently of all the others. Thus a system with n outputs can be treated, for purposes of determining weighting functions, as n independent systems, all having the same inputs.

Effect of Neglecting an Input

The results of the preceding discussion may be applied to show the effect of neglecting one input in the determination of the weighting functions of a system with two or more inputs. As an example, the effect of neglecting $y(t)$ in the system of Fig. 3 may be considered.

It is clear that if $x(t)$ and $y(t)$ are uncorrelated, the weighting function $g(t)$ can be determined from $x(t)$ and $x(t)$ alone, and no error can result from neglecting $y(t)$. Thus, if $y(t)$ is a noise disturbance uncorrelated with $x(t)$, it can have no effect on the determination of $g(t)$. However, if $x(t)$ and $y(t)$ are correlated, the weighting function that would be found from $x(t)$ and $x(t)$ alone, when $y(t)$ is neglected, would not be $g(t)$. Instead, it would be $g(t)$ plus the cascaded effect of $f(t)$ and $h(t)$. In terms of the equivalent causal path of Fig. 5, the weighting function found would be not the response for a unit impulse at the x-input to the system but rather the response for a unit impulse in $x(t)$ applied "upstream" of the equivalent causal path joining $x(t)$ and $y(t)$. This result can be generalized to systems with any number of inputs.

Nonlinear Systems

Since the weighting function is a linear representation of a system, the application of the techniques described in the foregoing section to a system that is in fact nonlinear still gives a linear representation. In this section it will be shown that this linear representation is the best linear representation of the system (in a mean-square sense) obtainable from the available data. To indicate the range of application of this linear representation in control problems, reference then will be made to a number of recent control-system studies using linear representations of nonlinear systems.

Since the equations of this paper are written for continuous input and output functions but for discrete weighting functions, two treatments of the problem are given. In the first treatment, given in this section, input and output functions as well as weighting functions are considered only at discrete points in time; this treatment is based on "least squares." In the second treatment, given in Appendix 1, weighting functions as well as input and output functions are treated as continuous; the solution then can be specialized to discrete points in time to give the equations used earlier. Both treatments are similar to the corresponding treatments which have been used to demonstrate methods for computing optimum linear filters (5). Throughout the discussion the assumption is made that the systems themselves are time invariant and that their inputs are statistically stationary; the implications of nonstationary inputs are discussed in Appendix 2.

Single-Input, Single-Output System. In the system represented by Fig. 1 the supposition is made that the system is now nonlinear, but that it is desired to represent $c(t)$ as closely as possible by a series of the form

$$\sum_{n=0}^{k} h_n m(t - nT)$$

As the standard of "as closely as possible," the condition is used that the mean-square error of approximation

$$M \equiv \overline{\left[c(t) - \sum h_n m(t - nT)\right]^2} \quad [28]$$

should be a minimum, where the bar denotes an average taken over all available data.

To choose the h_n so as to make M a minimum, a necessary condition is that

$$\frac{\partial M}{\partial h_i} = 0, \qquad i = 0, \ldots, k \quad [29]$$

Now

$$M \equiv \overline{c^2(t)} - 2 \sum_n h_n \overline{m(t - nT)c(t)} + \sum_n \sum_r h_n h_r \overline{m(t - nT)m(t - rT)}$$

hence

$$\frac{\partial M}{\partial h_i} = -2\,\overline{m(t - iT)c(t)} + 2 \sum_n h_n \overline{m(t - iT)m(t - nT)}$$

$$= -2\phi_{mc}(iT) + 2 \sum_n h_n \phi_{mm}(iT - nT)$$

Thus Equation [29] requires that

$$\phi_{mc}(iT) = \sum_n h_n \phi_{mm}(iT - nT), \qquad i = 0, \ldots, k \quad [30]$$

To show that Equation [30] actually makes M a minimum (rather than a maximum or a value that is neither maximum nor minimum), any small change in the h_n must be shown to result in a larger M. Therefore the effect of replacing the h_n of Equation [30] by $(h_n + \Delta h_n)$ is investigated. Equation [28] then becomes

$$M + \Delta M \equiv \overline{c^2(t)} - 2\sum_n (h_n + \Delta h_n)\,\overline{m(t-nT)c(t)} + \sum_n \sum_r (h_n + \Delta h_n)(h_r + \Delta h_r)\,\overline{m(t-nT)m(t-rT)}$$

with the result

$$\Delta M = -2\sum_n \Delta h_n\,\overline{m(t-nT)c(t)} + \sum_n \sum_r (h_n\Delta h_r + h_r\Delta h_n + \Delta h_r\Delta h_n)\,\overline{m(t-nT)m(t-rT)}$$

When the h_n are chosen to satisfy Equation [30], ΔM reduces to

$$\Delta M = \sum_n \sum_r \Delta h_r\Delta h_n\,\overline{m(t-nT)m(t-rT)} = \overline{\left[\sum_n \Delta h_n m(t-nT)\right]\left[\sum_r \Delta h_r m(t-rT)\right]} = \overline{\left[\sum_n \Delta h_n m(t-nT)\right]^2}$$

which is positive for all possible values of Δh_n as long as there is a nonzero input $m(t)$. Thus, since any change in the h_n from the values given by Equation [30] results in an increase in M, these values do make M a minimum. In the course of this proof the subscripts n and r have been interchanged freely since they are merely dummy subscripts of summation.

A comparison of Equation [30] with Equation [5] shows that while for a linear system Equation [5] should be fulfilled for all points of time, for a nonlinear system, Equation [5] should be fulfilled only at the values of τ corresponding to the values of t for which the weighting function is sought. This gives a $(k + 1)$ by $(k + 1)$ matrix equation which could be solved for the h_n by conventional methods. If the weighting function of a nonlinear system is found by using the DLS, then Equation [5] should be satisfied for the range

$$0 \leqq \tau \leqq kT$$

because the discrete points in Equation [29] are generally close enough together so that their individual identity need not be preserved.

The conclusions just reached should be compared with the conclusions reached in Appendix 1 of reference (3). There it was shown that when there is noise in a closed-loop system, Equation [5] should be satisfied in a region, $\tau \geqq A$, far enough to the right of the origin to eliminate the effect of the noise. Thus, in obtaining the weighting function of a nonlinear closed-loop system with noise, the engineer must use judgment in compromising between satisfying Equation [5] in the region of τ that gives the optimum linear representation of the system and satisfying Equation [5] in the region that minimizes the effects of noise in the closed loop.

Two-Input, Single-Output System. By application of the same procedure to the system represented by Fig. 3, the function to be minimized is now

$$M \equiv \overline{\left[x(t) - \sum g_n x(t-nT) - \sum h_n y(t-nT)\right]^2} \quad [31]$$

The choice of g_n and h_n to achieve this result requires that

$$\frac{\partial M}{\partial g_i} = 0 \text{ and } \frac{\partial M}{\partial h_i} = 0, \qquad i = 0, \ldots, k \quad [32]$$

Equating to zero the derivatives of M with respect to the g_i yields

$$\phi_{xx}(iT) = \sum_n [g_n\phi_{xx}(iT-nT) + h_n\phi_{xy}(iT-nT)] \quad [33]$$

and equating to zero the derivatives of M with respect to the h_i yields

$$\phi_{yx}(iT) = \sum_n [g_n\phi_{yx}(iT-nT) + h_n\phi_{yy}(iT-nT)] \quad [34]$$

To show that the values of g_n and h_n given by Equations [33] and [34] actually do make M a minimum, the same procedure used for Equation [28] can be followed.

Comparison of Equations [33] and [34] with Equations [7] and [8] reveals that the optimum linear representation of a two-input nonlinear system is obtained when Equations [7] and [8] are satisfied for the values of τ corresponding to the values of t for which the weighting function is sought.

By an extension of the reasoning of Appendix 1 of reference (3), it can be shown (6) that when there is noise in a closed-loop, two-input system, Equations [7] and [8] should be satisfied in a region, $\tau \geqq A$, far enough to the right of the origin to eliminate the effects of the noise. Thus, as in the case of a single-input system, a compromise must be made between satisfying these equations in the region of τ that gives the optimum linear representation and satisfying them in the region that minimizes the effects of noise.

The results of this section can be generalized to any number of inputs.

Relation Between Two Inputs. By the same type of argument, it can be shown that when two inputs x and y are not linearly related, the best linear time-series representation of y in terms of x (in a mean-square sense) is

$$y \approx \sum_{n=-k}^{k} f_n x(t-nT)$$

where the f_n are obtained by solving Equation [10] at the values of τ corresponding to the values of t for which the f_n are sought.

To show this, the expression

$$M \equiv \overline{[y(t) - \Sigma f_n x(t-nT)]^2}$$

is minimized by setting

$$\frac{\partial M}{\partial f_i} = 0, \qquad i = -k, \ldots, k$$

By expansion of M as before and by setting the derivatives equal to zero

$$\phi_{xy}(iT) = \sum_n f_n\phi_{xx}(iT-nT), \qquad i = -k, \ldots, k \quad [35]$$

and by the procedure used previously, it can be shown that these values of the f_n actually do make M a minimum.

Use of Linear Representations in Control Problems. While a linear representation admittedly does not give a complete description of a nonlinear system, it nevertheless gives a description that is often useful for control purposes. For many types of nonlinearities, a system becomes essentially linear when the input variations are infinitesimal, and essential information about the performance and stability of the system often can be determined from the linear model based on small perturbations (7). In records made under normal operating conditions, the random variations in inputs and outputs are likely to be so small that the weighting function determined from the normal operating records is the same as that given by small-perturbation theory.

When the random variations in normal operating records are so large that the essentially linear range of operation of the system is exceeded, a weighting function discovered from normal operating records may be compared to a "quasi linearization" of the system. Quasi linearization (8 to 11) is the replacement of each nonlinear

element in the system by an equivalent gain which is a function of the rms amplitude of the input; for each input-amplitude level, the equivalent gain is chosen to give the best approximation, in a mean-square sense, to the actual system output. When the input is sinusoidal, the quasi-linear representation of the output turns out to be its fundamental Fourier coefficient, and hence, quasi linearization is equivalent to discarding higher harmonics (8, 9). When the input is a random signal with a known probability distribution, the equivalent gain can be calculated from autocorrelations and cross-correlations (10) by use of a method similar to the method given earlier in this section. Both the quasi linearization based on sinusoidal inputs (8, 9, 11) and that based on random inputs (10) have proved useful in practice for systems that do not depart drastically from linearity; the approach based on random inputs has the advantage that random inputs are more realistic than sinusoidal inputs in an actual control situation.

The weighting functions determined from normal operating records correspond to the weighting functions for a quasi-linearized system with input-amplitude level equal to the actual level of the normal operating input. The use of normal operating records has the advantage that the weighting function is based on an averaging of data taken under actual operating conditions and hence is more realistic than a weighting function based on arbitrary input signals. A disadvantage of using normal operating records is that the data are taken only for the input-signal level occurring in the system; thus no information is obtained on the variation of quasi-linearized system parameters with signal level. This disadvantage sometimes can be offset if the system inputs have different signal levels under different operating conditions.

Much work remains to be done on the applicability of linearized representations to nonlinear control problems. The work just cited, however, indicates that linearized representations contain much useful information, especially when interpreted in the light of the known physical characteristics of the system.

To obtain further information about a nonlinear system from its normal operating records, the weighting function obtained by correlation techniques may be treated as a first approximation. The output for the linear approximation can be compared with the actual system output, and the difference between these two outputs can be used to determine the effects of nonlinearities under operating conditions. This procedure, suggested by Tustin (12), can be accomplished readily on the DLS. The weighting function is set up on the DLS, and the original system input is played through it; the DLS output then is subtracted from the original system output to give the portion of the output that is due to nonlinear effects.

Frequency Response

The frequency-response function, giving the attenuation and phase shift of a system for sinusoidal inputs of different frequencies, is often of interest. When written in complex-variable notation, the frequency-response function is the Fourier transform of the weighting function (1).

Three methods of obtaining the frequency-response function from normal operating records are:

1 Use the DLS to solve for the weighting function (Fig. 6); then experimentally measure the frequency response of the DLS, which serves as a model of the system. This method has been found to be simple and effective in practice and is the recommended method.

2 Obtain the weighting function by using the DLS or by numerical means; then numerically perform a Fourier transformation to obtain the frequency-response function. This method is not recommended if the weighting function is obtained on the DLS, because it does not take full advantage of the DLS.

3 Use a frequency analysis throughout. Instead of using correlation functions, their Fourier transforms, the spectral densities, may be used (1). Some simplification results from the fact that the Fourier transform of a convolution integral is a simple multiplication. Thus, by taking the Fourier transform of both sides of Equation [40] (see Appendix 1), which is the continuous form of Equation [5]

$$\Phi_{mc}(\omega) = H(j\omega)\,\Phi_{mm}(\omega) \quad \ldots\ldots\ldots\ldots [36]$$

from which the frequency-response function $H(j\omega)$ can be obtained merely by dividing the cross spectral density $\Phi_{mc}(\omega)$ by the input spectral density $\Phi_{mm}(\omega)$. This method appears attractive because of its evident mathematical simplicity; however, it has two pitfalls: (*a*) Since the frequency-response function does not show the directions of causal paths in a closed loop, there is no way of eliminating the effect of noise in a closed loop (see Appendix 1 of reference 3). (*b*) The process of taking the Fourier transform of the correlations is usually difficult in practice, especially if the correlations are based on short records of inputs and outputs. If, in spite of these difficulties, it is desired to use equations of the form of Equation [36], the frequency-response functions of a multi-input system also can be obtained by this means. The equations used are merely the Fourier transforms of Equations [7], [8], and [10] through [16] for the two-input case, and Equations [18] through [20] and [22] through [27] for the three-input case.

These Fourier transforms are valid only for statistically stationary inputs.

Conclusion

The method presented here should facilitate the determination of weighting functions of systems with multiple, mutually correlated inputs from normal operating records. For a nonlinear system, the optimum linear representation given by this method supplies useful information for control purposes, and may be used as a first step toward a more refined description of the system.

Acknowledgment

The benefit of discussions of this material with Prof. J. B. Reswick of Massachusetts Institute of Technology, and with Mr. J. M. Loeb of the Schlumberger Instrument Company, is greatly appreciated.

BIBLIOGRAPHY

1 "Theory of Servomechanisms," by H. M. James, N. B. Nichols, and R. S. Phillips, McGraw-Hill Book Company, Inc., New York, N. Y., 1947, chapter 2.

2 "Application of Statistical Methods to Communication Problems," by Y.-W. Lee, Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, Mass., September 1, 1950, Technical Report No. 181.

3 "Determination of System Characteristics From Normal Operating Records," by T. P. Goodman and J. B. Reswick, Trans. ASME, vol. 78, 1956, pp. 259–271.

4 "Determine System Dynamics Without Upset," by J. B. Reswick, *Control Engineering*, vol. 2, June, 1955, pp. 50–57.

5 "Extrapolation, Interpolation, and Smoothing of Stationary Time Series With Engineering Applications," by N. Wiener, The Technology Press, Massachusetts Institute of Technology, Cambridge, Mass., and John Wiley & Sons, Inc., New York, N. Y., 1949. (See especially Appendix B, "The Wiener RMS (Root Mean Square) Error Criterion in Filter Design and Prediction," and Appendix C, "A Heuristic Exposition of Wiener's Mathematical Theory of Prediction and Filtering," by N. Levinson.)

6 "Experimental Determination of System Characteristics From Correlation Measurements," by T. P. Goodman, ScD thesis, Department of Mechanical Engineering, Massachusetts Institute of Technology, Cambridge, Mass., 1955.

7 "Study of Pneumatic Processes in the Continuous Control of Motion With Compressed Air—Parts I and II," by J. L. Shearer, Trans. ASME, vol. 78, 1956, pp. 233–249.

8 "A Frequency Response Method for Analyzing and Synthesizing Contactor Servomechanisms," by R. J. Kochenburger, Trans. AIEE, vol. 69, part I, 1950, pp. 270–284.

9 "Sinusoidal Analysis of Feedback-Control Systems Containing Nonlinear Elements," by E. C. Johnson, Jr., Trans. AIEE, vol. 71, part II, 1952, pp. 169–181.

10 "Nonlinear Control Systems With Statistical Inputs," by R. C. Booton, Jr., Dynamic Analysis and Control Laboratory Report No. 61, Massachusetts Institute of Technology, Cambridge, Mass., March 1, 1952.

11 "Contributions to Hydraulic Controls—7, Analysis of the Effects of Nonlinearity in a Valve-Controlled Hydraulic Drive," by E. I. Reeves, Trans. ASME, vol. 79, 1957, pp. 427–432.

12 "The Mechanism of Economic Systems," by A. Tustin, William Heinemann, Ltd., London, England, 1955 Appendix.

Appendix 1

Optimum Continuous Weighting Functions for Nonlinear Systems

When the input and output of a system are given as continuous functions, the optimum linear weighting function may be found as a continuous function of time by using the calculus of variations.

Equation [2] may be made continuous in the weighting function h_n by passing to the limit as $T \to 0$ and $kT \to K$, giving the convolution integral

$$c(t) = \int_0^K h(\tau)m(t-\tau)d\tau$$

The upper limit of integration K must be large enough to include all of the significant portion of the weighting function; if necessary, it may be infinity.

For a nonlinear system, it is desired to approximate $c(t)$ as closely as possible by a convolution integral of this form. The problem therefore is to find the continuous function $h(\tau)$ that minimizes the expression

$$M \equiv \overline{\left[c(t) - \int_0^K h(\tau)m(t-\tau)d\tau\right]^2} \quad \ldots\ldots [37]$$

As before, the bar denotes an average taken over all available data. If the assumption is made that $h(\tau)$ represents a stable system (1), the convolution integral is finite for all bounded values of m, even when $K \to \infty$. It is therefore permissible to interchange the order of the operations of integration and averaging.

Since $h(\tau)$ is now a continuous function, rather than a sequence of parameters, the calculus of variations must be used to minimize M. Therefore, when $h(\tau)$ is replaced by $[h(\tau) + \Delta h(\tau)]$ in Equation [37]

$$M + \Delta M \equiv \overline{\left\{c(t) - \int_0^K [h(\tau) + \Delta h(\tau)]m(t-\tau)d\tau\right\}^2} \quad \ldots [38]$$

Here $\Delta h(\tau)$ is thought of as a small increment in $h(\tau)$ that is a continuous function of τ. By expansion of Equation [38] and subtraction of Equation [37], using μ as a second dummy variable of integration

$$\Delta M = -2\int \Delta h(\tau)\left\{\overline{m(t-\tau)c(t)} - \int h(\mu)\overline{m(t-\tau)m(t-\mu)}d\mu\right\}d\tau + \overline{\left[\int \Delta h(\tau)m(t-\tau)d\tau\right]^2} \quad \ldots\ldots [39]$$

A necessary condition to make M a minimum is that the sum of the terms involving the first power of $\Delta h(\tau)$ be zero for all possible $\Delta h(\tau)$. This condition is fulfilled if, and only if, the expression in the braces is zero throughout the range of integration; that is, if, and only if

$$\phi_{mc}(\tau) = \int_0^K h(\mu)\phi_{mm}(\tau-\mu)d\mu, \qquad 0 < \tau < K \ldots [40]$$

With this condition fulfilled, a sufficient condition to make M a minimum is that the sum of the terms involving the second power of $\Delta h(\tau)$ be positive for all nonzero $\Delta h(\tau)$. Since the last term of Equation [39] is positive for all nonzero $\Delta h(\tau)$, except for the trivial case $m(t) \equiv 0$, this condition is also fulfilled.

Comparison of Equation [40] with Equations [30] and [5] shows again that the optimum weighting function for a nonlinear system is obtained by matching the cross-correlation for the range of τ for which the weighting function is sought.

This result can be generalized to any number of input variables, and can be readily extended to prove Equation [35] in continuous form (6).

Appendix 2

Nonstationary Inputs

Many of the equations in this paper were simplified greatly by the assumption that the system inputs, and hence also the outputs, are statistically stationary; that is, that their statistical properties, including autocorrelations and cross-correlations, do not change with time. Inputs encountered in practice may not be stationary; even if they are stationary, the length of available input and output records may not be sufficient to smooth out all the instantaneous variations, with the result that correlations based on short records may be nonstationary.

Equations [1] through [10] are valid for both stationary and nonstationary inputs, while Equations [11] through [16] are valid only for stationary inputs. Similarly, Equations [17] through [23] are valid for both stationary and nonstationary inputs, while Equations [24] through [40] are valid only for stationary inputs.

The error introduced by applying an equation valid only for stationary inputs to an input which is nonstationary depends on the length of the records available for correlation in comparison to the length of the significantly nonzero part of the weighting functions. For example, if the available correlations are based on $(N + 1)$ ordinate values, spaced at a time interval T, while the nonzero parts of the weighting functions are of length kT, then the greatest error introduced in Equations [11] through [16] is the error of replacing

$$\overline{x(t-kT)x(t+\tau+kT)} \quad \text{by} \quad \overline{x(t)x(t+\tau+2kT)}$$

Thus the error involves at most only k terms of the $N + 1$ terms which are averaged to form the correlation functions; if $k = 20$ and $N = 1000$, the error may not be significant unless the inputs are drastically nonstationary, while if $k = 20$ and $N = 100$, the error may be appreciable. This illustrates the advantage of using a large number of ordinates in computing correlations, although this advantage always must be weighed against the greater cost in computing time.

For a two-input system with nonstationary inputs, Equations [11] through [16] may be used to give a first approximation to the solution; then Equations [7] and [8], which are valid for nonstationary inputs, may be used to refine the solution. A similar procedure may be used for a three-input system, since Equations [18], [19], and [20] are valid for nonstationary inputs.

To find the best linear representation of a nonlinear system with nonstationary inputs, it would be possible to rewrite Equations [28] through [35] and [37] through [40] in a much more complicated form that would make them applicable to nonstationary inputs. However, since the linear representation is at best only an approximation, such a procedure might not ever be worth while in practice.

Discussion

T. M. Stout.[4] A thorough job has been done in laying a theoretical basis for characterization of multi-input systems from normal operating records. If experience with the proposed techniques shows them to be accurate and convenient, control-system engineers may find them a useful addition to their tool kit.

The significance of this work in the nonlinear domain is less clear. It is well to remember that any quasi linearization of a nonlinear system has only restricted utility. The describing-function method (references 8 and 9 of the paper) provides a characterization for sinusoidal inputs. Ordinarily, but not necessarily, the describing function characterizes a system component which is only amplitude-dependent and which may be highly nonlinear; e.g., a relay. The describing-function method is concerned principally with absolute system stability but sheds some light on the relative stability or degree of damping.

The equivalent gain (reference 10 of the paper) provides a characterization for random signals and, again, is ordinarily determined for a system component which is amplitude but not frequency-dependent. The method provides a means, for example, for approximate determination of the mean-square error when the system is subjected to a random input.

The characterization discussed in this paper is a function of time which is also amplitude-dependent. (A recent paper[5] contains experimental data for a nonlinear servomechanism which show a considerable variation in the nature of the weighting function $h(t)$ with a change in the amplitude level of the random input signal.) If artificial manipulation of the input signal is prohibited, the amplitude dependence of $h(t)$ can be discovered only by waiting for a change in the input amplitude. In a chemical process, this delay could be a matter of days and characteristics of the plant itself might change before there was any change in the nature of the disturbances.

In addition, the region of application of the weighting function $h(t)$ is somewhat vague. It is not clear how $h(t)$ may be used for analysis of the system for which it was determined or of some larger system containing the component characterized by $h(t)$. The weighting function, of course, could be used to compute an output for the input from which it was originally determined; the result is an approximation to data already on hand and therefore not very valuable. Being a linear characterization determined for a single input amplitude level, it will give an approximation of indeterminate accuracy for inputs differing in amplitude level but otherwise similar. Calculation of step or sinusoidal response from $h(t)$ is impossible in principle, so results of any such calculations likewise would be of doubtful validity. Perhaps the situation may be summarized by saying $h(t)$ is useful only if the system is, in fact, so nearly linear that the term "nonlinear" is inappropriate.

Because of the potential importance of the techniques discussed in this paper, attention should be given to the problems involved in both determination and application of $h(t)$ which arise from the amplitude dependence of nonlinear phenomena. Possibly a considerable amount of additional work, leading to other papers on the subject, will be required to treat these questions adequately.

[4] The Ramo-Wooldridge Corporation, Los Angeles, Calif.

[5] "The Use of Correlation Techniques in the Study of Servomechanisms," by T. M. Burford, V. C. Rideout, and D. S. Sather, *Journal of the British Institution of Radio Engineers*, vol. 15, May, 1955, pp. 249–257.

Author's Closure

Mr. Stout has performed a useful service in describing more fully some of the limitations mentioned in the paper in connection with the application of the correlation-and-deconvolution technique to nonlinear systems. The author agrees that much additional work will be required to define these limitations adequately and to establish the range of useful application of this technique. This additional work will depend on the availability of data from industrial processes and on the co-operation of organizations that have access to such data.

For the present, it appears that the principal application to nonlinear systems of the correlation-and-deconvolution technique will be to obtain a linear model (the DLS set with the coefficients h_n) that can be used in analog studies of alternative schemes for controlling a system or of combinations of a system with other elements in a more complex control system. For a "mildly" nonlinear system, this model, representing as it does the major dynamic effects in the system, should prove sufficiently accurate for many useful control studies. For a "strongly" nonlinear system, it is conceivable that the method suggested by Tustin (reference 12 of the paper) can be used to find a correction to the linear model that can then be applied to devise a nonlinear model representing, to some extent, the amplitude dependence of the system. For a system whose normal operation cannot be disturbed, such a model would provide the best available information on the dynamic characteristics of the system.

Hunting Due to Lost Motion

By H. PORITSKY,[1] SCHENECTADY, N. Y.

Hunting of a servosystem, due to lost motion, say, in one of its mechanical links, but in absence of input signals, is considered. If the slack is assumed to be taken up suddenly, the motion is governed by a linear differential equation but with proper discontinuities when the direction of motion in the loose link is reversed. For the case of second-order systems it is shown that, if the characteristic roots are complex, a periodic hunting motion always exists, and that the system, no matter how it is started, will converge to this hunting motion. If the characteristic roots of the second-order system are real, then a periodic hunting motion exists, but depending upon how the system is started, it may converge to this motion or it may converge to a stable position at either end of the lost-motion band. Third and higher-order systems are studied in a similar way and the equations for determination of periodic hunting motion obtained. Second-order systems, in which the system "coasts" as the slack is taken up, are discussed briefly.

1 Introduction

The following is concerned with hunting of a servosystem, due to lost motion, say, in one of its mechanical links.

The particular system studied is shown schematically in Fig. 1. It is designed to make the output or load (motor angle) θ follow the input I. The error

$$\delta = I - \theta \quad \ldots\ldots [1]$$

is amplified by means of a hydraulic control to x, in the manner described by Equation [3], to produce an acceleration of the output or load $p^2\theta$ which is proportional to x, generated by the control.

It is assumed that lost motion of amount $\Delta x = 2h$ occurs in one of the mechanical links of the valve in the hydraulic control, as a result of which the motor torque, as well as $p^2\theta$, is proportional not to x but to $x - h$, $x + h$ according as x is increasing or decreasing.

We shall be concerned with the case of no input, $I = 0$, thus leading to Equation [4] for the error δ. Without lost motion, the system investigated is described by the third-order system of differential Equations [2] to [4]. With lost motion as indicated in Fig. 1, Equation [2] is replaced by Equation [5].

Since motion without input is investigated, the hunting motions, if any, are self-excited. However, some initial displacement away from the quiescent position is postulated, such as may exist after the signal I has been reduced to zero and the system presumably should settle down to rest. We do not consider the inaccuracies in follow-up for nonvanishing I due to lost motion.

In equation form the third-order system of Fig. 1, in absence of lost motion, is described by ($p = d/dt$)

$$p^2\theta = Ax \quad \ldots\ldots [2]$$

$$(p + B)x = (Cp + D)\delta \quad \ldots\ldots [3]$$

$$\theta + \delta = 0 \quad \ldots\ldots [4]$$

where A, B, C, D are positive constants. As just stated, the lost motion will be assumed to occur between x and θ; to include it one modifies Equation [2] in the manner indicated in Figs. 2, 3, and given by

$$p^2\theta = \left\{ \begin{array}{l} A(x - h) \text{ while } x \text{ is increasing} \\ A(x + h) \text{ while } x \text{ is decreasing} \end{array} \right\} \quad \ldots\ldots [5]$$

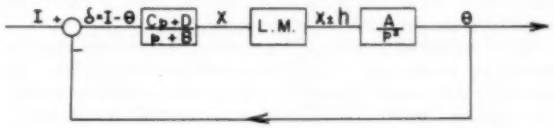


Fig. 1

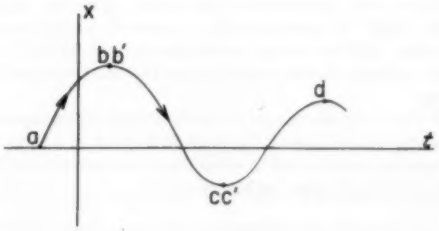


Fig. 2

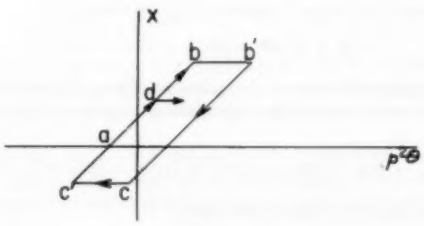


Fig. 3

Fig. 2 shows x versus t, while Fig. 3 shows a plot of $p^2\theta$ versus x. While x is increasing with time along the curve ab in Fig. 2, the relation between x and $p^2\theta$ is shown by the straight line ab in Fig. 3. When x starts decreasing with time, the representative point on Fig. 3 shifts suddenly from b to b', and as x decreases along bc in Fig. 2, the representative point in Fig. 3 describes the straight line $b'c$. Then, when in Fig. 2 x increases along cd, the line $c'd$ is followed in Fig. 3, and so forth. It is evident that in Fig. 3 the "hysteresis loops" are not of a fixed size, but have their horizontal sides at heights corresponding to the maxima and minima values of x.

---

[1] Consulting Engineer, General Electric Company. Mem. ASME.

Contributed by the Instruments and Regulators Division of The American Society of Mechanical Engineers and presented at the ASME-AIEE Conference on Nonlinear Control Systems, Princeton, N. J., March 26–28, 1956.

 Manuscript received at ASME Headquarters, January 19, 1956. Paper No. 56—IRD-12.

To start with, a simpler second-order system will first be considered, which differs from the preceding one in having $p^2\theta$ replaced by $p\theta$. This system is given by

$$(p + B)x + (Cp + D)\theta = 0 \quad \text{.....[6]}$$

$$p\theta = \left\{ \begin{array}{l} A(x - h) \text{ while } x \text{ is increasing} \\ A(x + h) \text{ while } x \text{ is decreasing} \end{array} \right\} \quad \text{.....[7]}$$

A corresponding change in replacing A/p^2 by A/p and in labeling the horizontal axis applies to Figs. 1 and 3.

It is shown that for the system given by Equations [6], [7], if the characteristic roots λ_1, λ_2 of the system are complex, a hunting or self-excited oscillation exists, and the system settles down to it, no matter how it is started. If the roots λ_1, λ_2 are real (and negative), i.e., if the system is overdamped, a hunting motion exists, but whether the system settles into it or settles to rest in the dead band depends upon how the system is started.

The present study dates from 1940–1941. The author wishes to thank the referees for calling his attention to the articles that have appeared in the literature in the meantime, some of which are listed as references at the end of the paper.

Many published treatments assume that the periodic hunting motion can be approximated by means of sinusoidal motion. Or else the first harmonic of the periodic hunting motion is obtained, and existing theory of linear systems is modified to include a transfer function for the lost motion; this function, however, will depend on the amplitude of the sinusoidal component. By contrast, the following does not resort to any such approximations, but utilizes exact solutions of the system equations, changing from one "orbit" to another as the x passes through a maximum or minimum. Of the several references one by Tustin (1)[3] comes closest to utilizing the same method as the one employed in the following.

While the treatment covers primarily the examples represented by the foregoing equations, the method is given in a form which can be applied equally well to other linear servosystems with lost motion, including systems of higher order.

2 Second-Order System—General Solution

We now consider the second-order system of Equations [6] and [7]. Suppose first that x is increasing. Applying $(Cp + D)$ to both sides of the proper Equation [7] there results

$$(Cp + D)p\theta = A(Cp + D)(x - h) \quad \text{.....[8]}$$

Interchanging the order of the operators $(Cp + D)$, p on the left-hand side and eliminating θ by means of Equation [6] there results

$$p(p + B)x + A(Cp + D)(x - h) = 0 \quad \text{.....[9]}$$

This also may be written in the form

$$\{\Delta(p) = [p(p + B) + A(Cp + D)]\}\ (x - h) = 0 \quad \text{..[10]}$$

since the derivatives of the constant $-h$ vanish. Thus while x is increasing, $x - h$ satisfies the second-order differential equation [10]. Similarly, it can be shown that $\theta + (hB/D)$ satisfies the same equation $\Delta(p) = 0$, while x is increasing. On the other hand, while x is decreasing, it can be shown that $x + h$ and $\theta - hB/D$ satisfy the same differential equation $\Delta(p) = 0$. At the transition point x and θ are continuous; however, the derivatives of x and θ are discontinuous there.

It will be convenient to introduce the variables

$$\xi = \left\{ \begin{array}{l} x - h, \\ \\ x + h, \end{array} \right. \qquad \eta = \left\{ \begin{array}{l} \theta + \dfrac{hB}{D} \\ \\ \theta - \dfrac{hB}{D} \end{array} \right. \quad \text{.....[11]}$$

where the upper or the lower definitions apply according to whether x is increasing or decreasing. Then it follows from the foregoing that, while x is either increasing or decreasing, ξ, η satisfy the homogeneous differential equation

$$\Delta(p) = p^2 + p(B + AC) + AD = 0 \quad \text{.....[12]}$$

as well as the homogeneous system of equations

$$\left. \begin{array}{l} p\eta = A\xi \\ (p + B)\xi + (Cp + D)\eta = 0 \end{array} \right\} \quad \text{.....[13]}$$

However, ξ, η undergo the sudden changes

$$\left. \begin{array}{l} \Delta\xi = \pm 2h \\ \Delta\eta = \mp \dfrac{2hB}{D} \end{array} \right\} \quad \text{.....[14]}$$

when x passes through extremal values, the upper signs applying when x passes through a maximum, the lower through a minimum.

In the following we shall consider the solution in terms of ξ and η, and plot it in the (ξ, η)-plane. We refer to the solutions of Equations [13] as "orbits." The (ξ, η) point will follow an orbit, say, while ξ is increasing; then it will jump to a new orbit in accordance with the upper signs in Equations [14]. The latter will be followed while ξ is decreasing, when the opposite $\Delta\xi$, $\Delta\eta$ jumps occur, etc.

It is readily shown that the solutions of Equations [13] may be expressed as linear combinations of two exponentials

$$\left. \begin{array}{l} \xi = A_1 e^{\lambda_1 t} + A_2 e^{\lambda_2 t} \\ \eta = A\left[\dfrac{A_1}{\lambda_1} e^{\lambda_1 t} + \dfrac{A_2}{\lambda_2} e^{\lambda_2 t}\right] \end{array} \right\} \quad \text{.....[15]}$$

where λ_1, λ_2 are the roots (assumed distinct) of the algebraic equation

$$\Delta(\lambda) = \lambda^2 + \lambda(B + AC) + AD = 0 \quad \text{.....[16]}$$

and the A_1, A_2 are arbitrary constants. Since the coefficients A, B, C, D are real, the roots of Equation [16] are either real or conjugate imaginaries. If the servosystem is not unstable, the real parts of the roots must be negative. This will be assumed to be the case.

If the constant A_2 in Equations [15] vanishes, then Equations [15] reduce to

$$\xi = A_1 e^{\lambda_1 t}, \quad \eta = \frac{AA_1}{\lambda_1} e^{\lambda_1 t} \quad \text{.....[17]}$$

Hence in the (ξ, η)-plane the representative point moves along the straight line

$$\eta - \frac{A}{\lambda_1}\xi = 0 \quad \text{.....[18]}$$

Similarly, if A_1 vanishes in Equations [15], the point moves along the straight line

$$\eta - \frac{A}{\lambda_2}\xi = 0 \quad \text{.....[19]}$$

[3] Numbers in parentheses refer to the Bibliography at the end of the paper.

If oblique Cartesian axes ξ_1, η_1 be introduced with these lines as their co-ordinate axes, by letting

$$\xi_1 = \xi - \frac{\lambda_2}{A}\eta, \quad \eta_1 = \xi - \frac{\lambda_1}{A}\eta \ldots\ldots\ldots [20]$$

then it is easy to show that the differential system given by Equations [13] transforms into

$$\frac{d\xi_1}{dt} = \lambda_1\xi_1, \quad \frac{d\eta_1}{dt} = \lambda_2\eta_1 \ldots\ldots\ldots\ldots [21]$$

Its solution is given by

$$\xi_1 = B_1e^{\lambda_1 t}, \quad \eta_1 = B_2e^{\lambda_2 t} \ldots\ldots\ldots\ldots [22]$$

where B_1, B_2 are arbitrary constants, related to A_1, A_2 as follows

$$B_1 = A_1\left(1 - \frac{\lambda_2}{\lambda_1}\right), \quad B_2 = A_2\left(1 - \frac{\lambda_1}{\lambda_2}\right) \ldots\ldots [23]$$

The variables ξ_1, η_1 which cause the system to reduce to the simple form given by Equations [21] in which the variables separate, constitute the "normal co-ordinates" of the system.

3 Real Roots

If λ_1, λ_2 are real, then they are both negative, say as shown in Fig. 4; we assumed $\lambda_1 \neq \lambda_2$. The orbits in the (ξ_1, η_1)-plane described by Equations [22] are then as shown in Fig. 5. Except

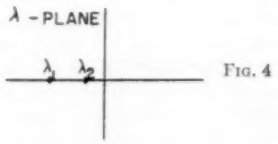


Fig. 4

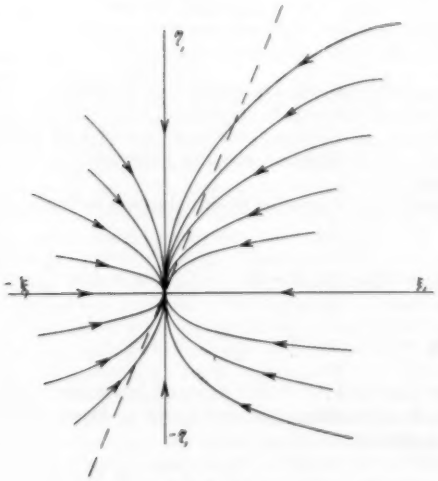


Fig. 5

for the two axes, they are fractional power parabolas, given by

$$\eta_1 = \text{const}\ |\xi_1|^{\lambda_2/\lambda_1} \ldots\ldots\ldots\ldots [24]$$

which equation is obtained from Equations [22] by eliminating t. All the curves given by Equations [22] are tangent to the η_1-axis, with exception of the two orbits constituting the positive and negative ξ_1-axis.

The relation between ξ, η and ξ_1, η_1 given by Equations [20] is called an "affine linear transformation." Such a transformation between two planes changes parallel lines into parallel lines, but in general conserves neither distances nor angles. For simplicity we may refer to it as a "skewing" transformation. Solving Equations [20] for ξ, η one obtains for the inverse a similar affine or skewing transformation. Hence the orbits in the (ξ, η)-plane are as shown in Fig. 6, and form a skew image of the orbits of Fig. 5.

The two lines given by Equations [18] and [19] correspond, respectively, to the η_1, ξ_1-axes of Fig. 5, in view of Equations [20]. In Fig. 6, since λ_1, λ_2 are negative, these lines lie in the second and

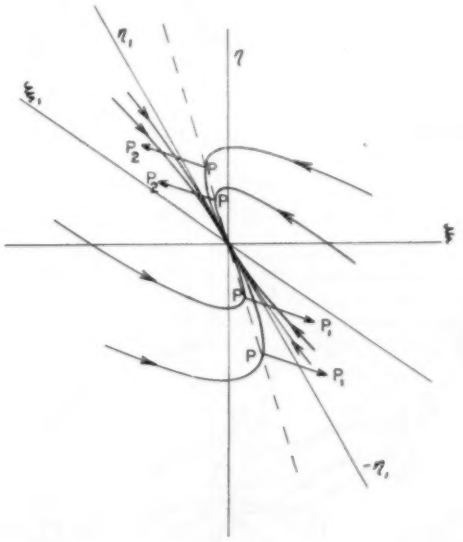


Fig. 6

fourth quadrants. Except for the ξ_1-axis, the orbits are tangent to the η_1-axis.

By eliminating dt from Equations [13] one obtains for the reciprocal slope of the orbits

$$\frac{d\xi}{d\eta} = \frac{d\xi/dt}{d\eta/dt} = -\left(C + \frac{B}{A} + \frac{D}{A}\frac{\eta}{\xi}\right) \ldots\ldots [25]$$

This slope is a function of η/ξ only, and hence is the same at all points of the straight lines through the origin. Hence, or directly from Equations [15], follows that the orbits are similar about the origin.

It will now be recalled that in the (ξ, η)-plane an orbit is followed until the sign of $d\xi/dt$ starts changing. From Equation [25] it will be seen that this occurs along the straight line

$$\frac{\eta}{\xi} = -\left(\frac{AC + B}{D}\right) \ldots\ldots\ldots\ldots [26]$$

This is shown as a dashed line in Figs. 5, and 6. When the orbit hits the lower half of this line, a sudden shift such as PP_1 corresponding to the upper signs of Equations [14] takes place; when it hits the upper half of this line, a similar shift PP_2 corresponding to the lower signs of Equations [14] occurs.

Following through the various orbits and jumps, we arrive at the possibilities shown on Fig. 7. If after a jump the point lands in the regions with slanted shading, no further change of sign of $d\xi/dt$ occurs, and the point moves in toward the origin along a proper orbit. This happens along the orbits in the vertically shaded regions of Fig. 7. The orbits in these regions are the orbits

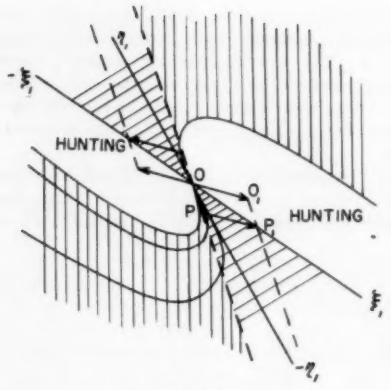


FIG. 7

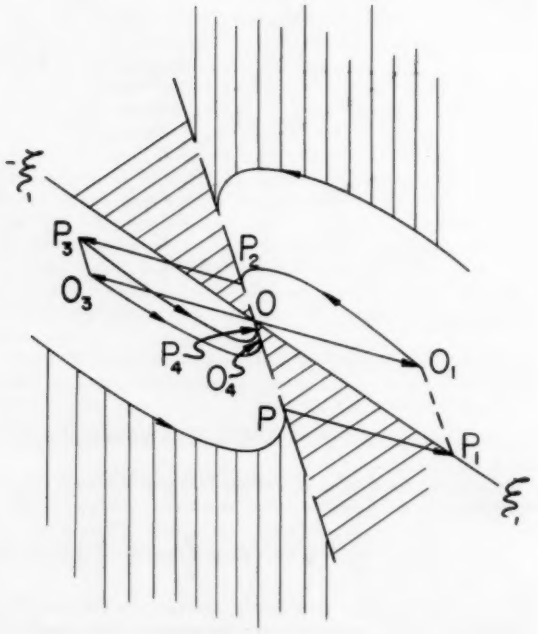


FIG. 8

which after one translation become stable. Between the shaded areas are regions where hunting takes place. The hunting motions approach a permanent oscillation of fixed amplitude.

To prove this consider Fig. 8 which is a magnified image of the part of Fig. 7 near the origin. It will be noted that the orbits from the unshaded region below the negative ξ_1-axis hit the dashed line between O and P. They are then translated to O_1P_1, and, following the curves of the new region, hit the dashed line again at OP_2, whence they are translated into P_2O_2. Thus they have been compressed to a small region of the original unshaded area, and, in fact, they hit the dotted line again between O_4P_4 over a much smaller interval than OP. This compression continues as the motion proceeds, with x increasing and decreasing alternately, the width of the corresponding interval decreasing eventually in geometric ratio. The various motions, therefore, approach a certain limiting curve which is the present oscillation or hunting motion in question.

The amplitude of the permanent oscillation is proportional to h, and likewise the width of the region from which the permanent oscillation develops is proportional to h.

While so far the motion has been described in the (ξ, η)-plane of Figs. 7, 8, it is of interest also to describe it in the original (x, θ)-plane. Recalling Equations [11] it will be found that the upper signs apply to the left of the dashed line $P_2O \ldots P$ of Fig. 8, where x is increasing along the orbits. This half plane then suffers a translation relative to the (ξ, η)-plane of Fig. 8 so that the origin O is moved to O_1. Similarly, the half plane to the right of this dashed line is translated so that O is moved to O_2. The motion in the (x, θ)-plane can therefore be described in terms of skewed images of fractional-power parabolas with vertexes at two different points, O_1, O_2. There is the advantage that the jumps are eliminated. On the other hand, the orbits converging to O_1 and O_2 overlap each other. The "stable" motions converge to O_1 or O_2.

It is evident that the periodic hunting motion is by no means an ellipse, which it would be if x and θ were sinusoidal in time.

4 Complex Roots

If λ_1, λ_2 are complex, then as shown in Fig. 9, they are given by

$$\lambda_1 = \mu + i\nu, \quad \lambda_2 = \mu - i\nu \ldots\ldots\ldots\ldots [27]$$

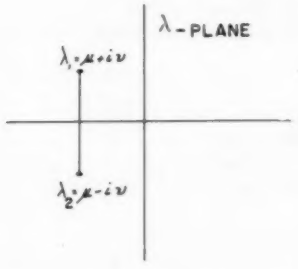


FIG. 9

where μ, ν are real and μ is negative. Equations [15 to 23] and the transformation to ξ_1, η_1 are still valid, but for real ξ, η lead to complex ξ_1, η_1 which are conjugate complex of each other. Similarly, the straight lines given by Equations [18, 19] turn out imaginary.

This difficulty is remedied by introducing a further transformation

$$\left.\begin{aligned} \xi_2 &= (\xi_1 + \eta_1)/2 = \xi - \frac{\lambda_1 + \lambda_2}{2A}\eta = \xi - \frac{\mu}{A}\eta \\ \eta_2 &= (\xi_1 - \eta_1)/2i = \frac{\lambda_1 - \lambda_2}{2Ai}\eta = \frac{\nu}{A}\eta \end{aligned}\right\} \ldots [28]$$

Since in Equations [15] A_1 and A_2 are conjugate imaginaries, so are also B_1, B_2 in Equations [22], and ξ_1 and η_1; hence ξ_2, η_2 are real and are given by

$$\left.\begin{aligned} \xi_2 &= R(\xi_1) = R\left[A_1\left(1 - \frac{\lambda_2}{\lambda_1}\right)e^{\lambda_1 t}\right] \\ \eta_2 &= I(\xi_1) = I\left[A_1\left(1 - \frac{\lambda_2}{\lambda_1}\right)e^{\lambda_1 t}\right] \end{aligned}\right\} \ldots\ldots [29]$$

where $R(z)$, $I(z)$ denote, respectively, the real and imaginary parts of z.

From Equations [29] it follows that the points (ξ_2, η_2) describe a logarithmic spiral in the (ξ_2, η_2)-plane as shown in Fig. 10.

Equations [28] show that between the (ξ, η)-plane and the (ξ_2, η_2)-plane there exists a relation of the same kind as between the

planes of Fig. 5 and Fig. 6, namely, a skewing or linear homogeneous transformation, sending parallel lines into parallel lines, but deforming circles into ellipses. Thus in the original (ξ, η)-plane the orbits are skewed images of logarithmic spirals.

Equations [25] and [26] still apply here. The points at which ξ attains its maxima or minima are once more given by the straight line Equation [26] in the (ξ, η)-plane; in view of the linear relation between the variables ξ, η and ξ_2, η_2, this straight line transforms into a proper straight line through the origin in the (ξ_2, η_2)-plane shown as a dashed line aOb in Fig. 10. It is now

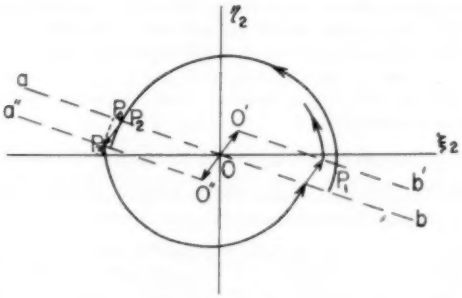


FIG. 10

clear that all orbits, no matter where they start, necessarily intersect aOb eventually. Thus if ξ is increasing, it will necessarily attain a maximum and start decreasing again. The maximum is taken on when the orbit meets the right half Ob of the dashed line; then the values of ξ_2 and η_2 are shifted by a constant vector which corresponds to upper signs in Equations [14]. Starting along new logarithmic spirals from the displaced points, along $O'b'$, one continues along an orbit until it cuts the dashed straight line along aO, whereupon a translation in the opposite direction, to $a''O''$, takes place, and so forth.

It will be noted that the logarithmic spiral which starts far away from the origin converges toward it appreciably during one of these "half cycles." On the other hand, a point starting very close to the origin gets farther away from it, largely on account of the fixed translation. Thus there is one point on the dashed line Ob which after the translation and convergence along the spiral, finds itself at the same distance from the origin along ab. This point describes a periodic motion. The curves that start on the outside of this closed path get smaller and converge toward it, while the curves that start on its inside expand and approach it.

Thus for the case of complex characteristic roots there always exists a self-oscillation or hunting due to the lost motion, and no matter where the system is started, it necessarily approaches this self-oscillation as a limiting motion.

As in Section 3 the motion may be described in the original (x, θ)-plane. Here the orbits are continuous and form segments of skewed images of equiangular spirals converging to two different centers. For small μ/ν they more nearly resemble ellipses.

While the foregoing treatment is largely descriptive and geometric, a purely analytic treatment can also be given. This is done in Sections 5, 6 for third-order and nth-order systems; the latter includes the case $n = 2$.

5 Third-Order System—Real Roots

We consider next the third-order system given by Equations [3 to 5], and possessing lost motion as represented by Equation [5] and Fig. 3. As x passes through a maximum or minimum, x, θ, px, $p\theta$ are continuous, but $p^2\theta$ is discontinuous in accordance with Equation [5].

We introduce the variables

$$\xi = x \mp h, \quad \eta = \theta \pm hB/D, \quad \zeta = p\theta \ldots\ldots [30]$$

where the upper or lower signs apply accordingly as x is increasing or decreasing. It is readily shown that between maxima and minima of x, the variables ξ, η, ζ satisfy the homogeneous system

$$(p + B)\xi + (Cp + D)\eta = 0, \quad p\eta = \zeta, \quad p\zeta = A\xi \ldots [31]$$

while each one is a solution of

$$\Delta(p) = p^3 + Bp^2 + ACp + AD = 0 \ldots\ldots [32]$$

When x passes through an extremal value, the variables ξ, η, ζ change as follows

$$\Delta\xi = \pm 2h, \quad \Delta\eta = \mp 2hB/D, \quad \Delta\zeta = 0 \ldots\ldots [33]$$

with the upper signs corresponding to passage of x through a maximum.

In general the characteristic equation

$$\Delta(\lambda) = \lambda^3 + B\lambda^2 + AC\lambda + AD = 0 \ldots\ldots [34]$$

of Equations [31] or [32] will have three unequal roots, $\lambda_1, \lambda_2, \lambda_3$, and the solutions for ξ, η, ζ will reduce to linear combinations of the exponentials $e^{\lambda_1 t}, e^{\lambda_2 t}, e^{\lambda_3 t}$. The roots λ_i are either all real and negative as in Fig. 11, or else one, say λ_3, is real and negative, and two are conjugate complex, with a negative real part, as in Figs. 12(a, b).

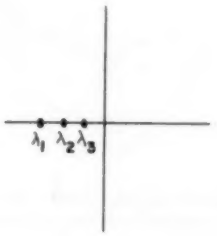
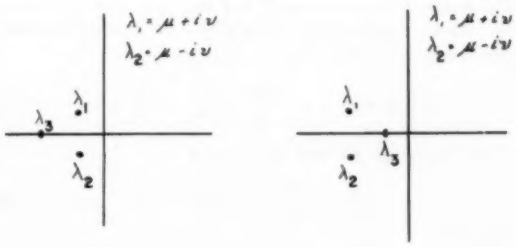


FIG. 11



(a) (b)

FIG. 12

When the roots are real, it is possible to introduce normal co-ordinates ξ_1, η_1, ζ_1 as proper linear combinations of ξ, η, ζ, for which the system given by Equations [31] reduces to

$$\frac{d\xi_1}{dt} = \lambda_1\xi_1, \quad \frac{d\eta_1}{dt} = \lambda_2\eta_1, \quad \frac{d\zeta_1}{dt} = \lambda_3\zeta_1 \ldots\ldots [35]$$

and whose solutions are therefore given by

$$\xi_1 = A_1e^{\lambda_1 t}, \quad \eta_1 = A_2e^{\lambda_2 t}, \quad \zeta_1 = A_3e^{\lambda_3 t} \ldots\ldots [36]$$

This is done by seeking solutions of Equations [31] of the form

$$\xi = B_1e^{\lambda t}, \quad \eta = B_2e^{\lambda t}, \quad \zeta = B_3e^{\lambda t} \ldots\ldots [37]$$

where B_i, λ are constants. Substitution of Equations [37] into Equations [31] leads to three homogeneous linear equations in B_i, which admit solutions other than $B_1 = B_2 = B_3 = 0$ if and only if the determinant vanishes. In this way Equation [34] results. Corresponding to each root of Equation [34] (assumed real and distinct) the three equations in B_i become linearly dependent and of rank 2, and determine unique ratios $B_1:B_2:B_3$. Equations [37] then yield a straight-line solution for each root, and these lines form the axes of a set of oblique Cartesian co-ordinates which are the normal variables ξ_1, η_1, ζ_1. The positive direction and the scale on each axis may be chosen so that Equation [38] in the following is valid.

It is readily shown that the solutions possess similarity about the origin of the (ξ, η, ζ) or (ξ_1, η_1, ζ_1)-space, such that if a conical surface with the vertex at the origin be drawn through any one orbit C_0, the curves on the same conical surface obtained by magnifying C_0 by any factor also will be orbits. Therefore the orbits may be followed to a certain extent by investigating the intersection of these conical surfaces with a (unit) sphere Σ with center at the origin. Such a sphere is shown on Fig. 13 where the great

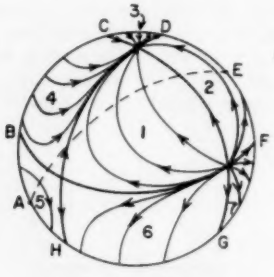


FIG. 13

circles correspond to orbits lying in one of the co-ordinate planes. The projection of a general orbit on each co-ordinate plane is a fractional power parabola. All the orbits approach the origin—this is not indicated in Fig. 13.

The linear relations between ξ, η, ζ and ξ_1, η_1, ζ_1 may be solved for ξ. By multiplying ξ_1, η_1, ζ_1, by proper constants, if necessary, ξ may be expressed as follows

$$\xi = x \pm h = \frac{\xi_1}{\lambda_1} + \frac{\eta_1}{\lambda_2} + \frac{\zeta_1}{\lambda_3} \quad \ldots\ldots\ldots [38]$$

From Equations [35] and [38] follows

$$\frac{dx}{dt} = \frac{d\xi}{dt} = \xi_1 + \eta_1 + \zeta_1 \quad \ldots\ldots\ldots [39]$$

$$\frac{d^2x}{dt^2} = \lambda_1\xi_1 + \lambda_2\eta_1 + \lambda_3\zeta_1 \quad \ldots\ldots\ldots [40]$$

Now at a maximum of x Equations [39], [40] yield

$$\Pi_1: \xi_1 + \eta_1 + \zeta_1 = 0,\ \lambda_1\xi_1 + \lambda_2\eta_1 + \lambda_3\zeta_1 < 0 \ldots [41]$$

while at a minimum

$$\Pi_2: \xi_1 + \eta_1 + \zeta_1 = 0,\ \lambda_1\xi_1 + \lambda_2\eta_1 + \lambda_3\zeta_1 > 0 \ldots [42]$$

The loci Π_1, Π_2 constitute two halves of the plane

$$\Pi: \xi_1 + \eta_1 + \zeta_1 = 0 \quad \ldots\ldots\ldots [43]$$

divided by the straight line which forms the intersections of Π with the plane

$$\lambda_1\xi_1 + \lambda_2\eta_1 + \lambda_3\zeta_1 = 0 \quad \ldots\ldots\ldots [44]$$

From the foregoing it follows that an orbit is followed till it hits either Π_1 or Π_2 and then the changes given by Equations [33] take place, or in terms of the normal variables

$$\Delta\xi_1 = \pm a, \quad \Delta\eta_1 = \pm b, \quad \Delta\zeta_1 = \pm c \quad \ldots\ldots [45]$$

where a, b, c are proper constants obtained by expressing ξ_1, η_1, ζ_1 in terms of ξ, η, ζ and forming similar combinations of the right-hand members of Equation [33]. The upper signs of Equations [45] correspond, say, to an intersection of the orbit with Π_1, the lower signs with Π_2.

In Fig. 13 the plane Π is shown as the great circle $ABC \ldots HA$ forming the boundary of the visible half of Σ; the plane given by Equation [44] as the dashed half circle AE; the half planes Π_1, Π_2 as the 180-deg arcs $ABCDE$, $EFGHA$, respectively.

Suppose that an orbit has reached Π_1 at P_0:$(\xi_{10}, \eta_{10}, \zeta_{10})$ satisfying Equation [41], say at the time $t = 0$. Equations [45] show that it is displaced to

$$P_0'\text{:}(\xi_{10} + a, \quad \eta_{10} + b, \quad \zeta_{10} + c) \quad \ldots\ldots [46]$$

It now follows from Equations [36] that the resulting motion will continue along an orbit through P_0'

$$\left.\begin{aligned} \xi_1 &= (\xi_{10} + a)e^{\lambda_1 t} \\ \eta_1 &= (\eta_{10} + b)e^{\lambda_2 t} \\ \zeta_1 &= (\zeta_{10} + c)e^{\lambda_3 t} \end{aligned}\right\} \quad \ldots\ldots\ldots [47]$$

until it intersects Π_2 at P_1, or if it never cuts Π_2, forever converging to the origin.

The class of orbits cutting Π_2 can be obtained from those cutting Π_1 by negative reflection through the origin O. It follows that a periodic motion will result if P_1 is the negative image of P_0, or dropping the subscript $_0$ if for any ξ_1, η_1, ζ_1 satisfying Equation [43] the following equations hold for $t \geqq 0$

$$\left.\begin{aligned} (\xi_1 + a)e^{\lambda_1 t} &= -\xi_1 \\ (\eta_1 + b)e^{\lambda_2 t} &= -\eta_1 \\ (\zeta_1 + c)e^{\lambda_3 t} &= -\zeta_1 \end{aligned}\right\} \quad \ldots\ldots\ldots [48]$$

If Equations [48], [43] are satisfied for the same ξ_1, η_1, ζ_1 for several positive t, the smallest root is to be used.

Solving Equations [48] for ξ_1, η_1, ζ_1 and substituting in Equation [43] there results

$$\frac{a}{1 + e^{-\lambda_1 t}} + \frac{b}{1 + e^{-\lambda_2 t}} + \frac{c}{1 + e^{-\lambda_3 t}} = 0 \quad \ldots\ldots [49]$$

At $t = 0$ the left-hand member of Equation [49] reduces to $(a + b + c)/2$; for $t \to \infty$, essentially to $ce^{\lambda_3 t}$. Hence if $a + b + c$, c are of opposite sign, roots of Equation [49] will exist. Thus, at least for negative $1 + (a + b)/c$, there will exist periodic hunting motions.

Summarizing, it has been shown that for real λ_i a periodic hunting solution exists if Equation [49] has a positive root.

If Equation [49] has positive roots, the first positive root $t = t_1$ is substituted in Equations [48] to yield a point P_0 on Π_1

$$\left.\begin{aligned} P_0\text{:}[\xi_1 &= \xi_{10} = -a/(1 + e^{-\lambda_1 t_1}), \\ \eta_1 &= \eta_{10} = -b/(1 + e^{-\lambda_2 t_1}), \\ \zeta_1 &= \zeta_{10} = -c/(1 + e^{-\lambda_3 t_1})] \end{aligned}\right\} \quad \ldots\ldots [50]$$

Half the periodic orbit is then given by Equations [36] with $A_1 = \xi_{10}$, $A_2 = \eta_{10}$, $A_3 = \zeta_{10}$ for $-t_1 < t < 0$, while the other half is its negative image in the origin.

Some motions will approach the periodic motion; others will converge to the origin.

In the (ξ, η, ζ)-space the orbits form a skew image of the orbits of the (ξ_1, η_1, ζ_1)-space described by Equations [36] and the same statement applies to the translations following intersections with the skew images of Π_1, Π_2, and to the periodic motions.

In the (x, θ, ζ)-space the orbits are obtained by effecting opposite translation of the two half spaces to each side of image of Π.

It is clear that a similar treatment can be applied to a servo-system represented by a set of linear differential equations of order n, and subject to lost motion, at least for the case when the characteristic roots $\lambda_1, \ldots, \lambda_n$ are all real and distinct. One introduces normal co-ordinates $\xi_1, \ldots \xi_n$ and calculates their increments $\pm a_1, \ldots \pm a_n$ corresponding to "taking up the slack." The determination of the hunting motion can then be reduced to the determination of the first positive root of the equation

$$\sum_{i=1}^{n} \frac{a_i}{1 + e^{-\lambda_i t}} = 0 \quad \ldots\ldots\ldots\ldots\ldots\ldots [51]$$

This, of course, includes the case $n = 2$ of Section 3; for this case the left side of Equation [51] reduces to a sum of two terms.

6 Third-Order System—Complex Roots

If the three roots of Equation [34] are not all real, then they are as shown in Figs. 12(a, b). We proceed to discuss these cases.

As in the case of real roots, it is possible to introduce normal co-ordinates by means of which the system given by Equations [31] is reduced to the form of Equations [35]. However, with λ_1, λ_2 complex, as in Equations [27], ξ_1, η_1 will also be conjugate complex for real solutions. One may avoid this difficulty, as in Section 4, by introducing another set of variables

$$\xi_2 = (\xi_1 + \eta_1)/2, \quad \eta_2 = (\xi_1 - \eta_1)/2i, \quad \zeta_2 = \zeta_1 \ldots [52]$$

in terms of which the differential equations become real. Equations [52] yield

$$\xi_1 = \xi_2 + i\eta_2, \quad \eta_1 = \xi_2 - i\eta_2, \quad \zeta_2 = \zeta_1 \ldots\ldots [53]$$

Equations [53] show that one may use up the whole (ξ_2, η_2)-plane as the complex Gaussian plane for the complex variable ξ_1. Since η_1 is the conjugate of ξ_1 no separate representation for it is needed.

Expressing ξ_1 in polar co-ordinates

$$\xi_1 = re^{i\epsilon} \quad \ldots\ldots\ldots\ldots\ldots\ldots [54]$$

it follows from the Solution [36] that

$$\left.\begin{aligned} r &= |A_1|e^{\mu t} \\ \epsilon &= \arg A_1 + t\nu \\ \zeta_2 &= A_2 e^{\lambda_2 t} \end{aligned}\right\} \quad \ldots\ldots\ldots\ldots [55]$$

Equations [55] show that the projection of the orbit on the plane $\zeta_2 = 0$ describes a convergent logarithmic spiral, while its distance from this plane decreases exponentially.

It is evident that in the (ξ_2, η_2, ζ_2)-space the conical surfaces on which the orbits lie will now have an appearance quite different from that of Fig. 13 and will wind about the ζ_2-axis.

Elimination of t between r and ζ_2 shows that the orbits lie on surfaces of rotation

$$\zeta_2 = \text{const } (r^{\lambda_2/\mu}) \quad \ldots\ldots\ldots\ldots\ldots [56]$$

whose sections by planes $\epsilon = \text{const}$ are fractional power parabolas. If $\lambda_2 > \mu$, these surfaces are tangent to the ζ_2-axis at the origin O as in Fig. 14; if $\lambda_2 < \mu$ they are normal to the ζ_2-axis at O, as shown in Fig. 15.

By multiplying ξ_1, η_1, ζ_1, by proper constants if necessary, one may represent ξ in the form

$$\xi = R(\xi_1/\lambda_1) + \zeta_1/\lambda_2 \quad \ldots\ldots\ldots\ldots\ldots [57]$$

As in Section 5, one now describes the motion by following the orbits described by Equations [55] with x and ξ increasing until the orbit either converges to the origin or cuts the locus

$$R(\xi_1) + \zeta_1 = 0, \quad R(\xi_1\lambda_1) + \lambda_2\zeta_1 < 0; \quad \Pi_1 \ldots\ldots [58]$$

whereupon a translation corresponding to the upper signs of

$$\Delta\xi_2 = \pm a, \quad \Delta\eta_2 = \pm b, \quad \Delta\zeta_2 = \pm c \ldots\ldots [59]$$

takes place where a, b, c, are the changes in ξ_2, η_2, ζ_2 corresponding to the upper signs in Equations [33]. The motion is then followed with ξ decreasing until it converges to the origin or intersects the locus

$$\xi_2 + \zeta_2 = 0, \quad R(\xi_1\lambda_1) + \lambda_2\zeta_2 > 0; \quad \Pi_2 \ldots\ldots [60]$$

when a translation with the lower signs of Equations [59] takes place.

The left-hand member of the inequalities in Equations [58], [60], can be put in the form

$$\xi_2\mu - \eta_2\nu + \lambda_2\zeta_1 \quad \ldots\ldots\ldots\ldots\ldots [61]$$

and vanishes along a plane, thus showing that Π_1, Π_2 are half planes.

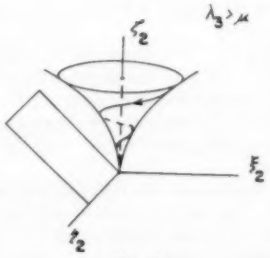


Fig. 14

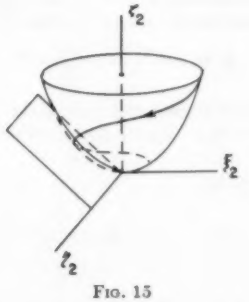


Fig. 15

The search of periodic motions leads to the equations

$$\left.\begin{aligned} (\xi_1 + a + ib)e^{\lambda_1 t} &= -\xi_1 \\ (\zeta_2 + c)e^{\lambda_2 t} &= -\zeta_2 \end{aligned}\right\} \quad \ldots\ldots\ldots [62]$$

for ξ_1, ζ_2 satisfying Equation [64] which follows. Solving for ξ_1, ζ_2 we obtain

$$\xi_1 = -\frac{a + ib}{e^{-\lambda_1 t} + 1}, \quad \zeta_2 = -\frac{c}{e^{-\lambda_2 t} + 1} \quad \ldots\ldots [63]$$

Substituting in

$$R(\xi_1) + \zeta_2 = \xi_2 + \zeta_2 = 0 \quad \ldots\ldots\ldots\ldots [64]$$

there results

$$\frac{c}{e^{-\lambda_3 t}+1}+R\left(\frac{a+ib}{e^{-\lambda_1 t}+1}\right)=0 \quad \ldots\ldots\ldots [65]$$

For $t = 0$, the left-hand member of Equation [65] reduces to $(c + a)/2$. For large t it becomes essentially

$$ce^{\lambda_3 t}+R[(a+ib)e^{\lambda_1 t}] = ce^{\lambda_3 t}+e^{\mu t}[a\cos \nu t - b \sin \nu t] \quad \ldots\ldots [66]$$

If $\lambda_3 < \mu$ as in Fig. 12(a), the last term in Equation [66] predominates as t increases and its oscillatory nature shows that roots of Equation [65] will exist. If $\lambda_3 > \mu$ as in Fig. 12(b), then the term $ce^{\lambda_3 t}$ predominates for large t and it may happen that no roots of Equation [65] exist. A more detailed examination of Equation [65] is now required. Certainly if $a + c$, c are opposite in sign roots will exist.

The two cases are shown in Figs. 14 and 15. Fig. 15 corresponds to $\lambda_3 < \mu$ as in Fig. 12(a). Here the surface given by Equation [56] on which the orbits lie always cuts the oblique plane Equation [64] as ζ_3 approaches zero, and roots of Equation [65] will exist. Fig. 14 corresponds to $\lambda_3 > \mu$ as in Fig. 12(b). Now the surface Equation [56] and the orbits on it, once they are above the plane, Equation [64], avoid it as ζ_3 approaches zero. Only if the translation given by Equations [63] transfers the point from one side of Equation [64] and $\zeta_3 = 0$ to the other side is a periodic hunting motion possible.

For the case of Fig. 14, it is evident that once an orbit subtends an angle less than 45 deg with the positive (or negative) ζ_3-axis it will approach the origin without ever cutting the oblique plane.

No detailed study of the regions of space analogous to those of Figs. 7 and 8, constituting the "stable" and the "hunting" regions, has been made.

Summarizing, it has been shown that hunting periodic motions certainly exist if $\lambda_3 < \mu$ and may exist even if $\lambda_3 > \mu$, though in the latter case only for limited initial conditions will the motion approach these periodic hunting motions. Half of the periodic motion is obtained by finding the first positive root $t = t_1$ of Equation [66], substituting it in Equations [62] to solve for ξ_1, $\zeta_3 = \zeta_{31}$, and substituting the resulting values $\xi_1 = \xi_{10}$, $\zeta_3 = \zeta_{30}$ in

$$\xi_1 = \xi_{10}e^{\lambda_1 t}, \quad \zeta_3 = \zeta_{30}e^{\lambda_3 t} \quad \ldots\ldots\ldots\ldots [67]$$

for $-t_1 < t < 0$. The other half is the negative image in the origin of the foregoing half.

While the foregoing equations appear to differ considerably from those of the preceding section, actually they can be made to agree with them in form. Equation [65] can be replaced by Equation [51] for $n = 3$, provided a_1, a_2, a_3 denote the increments in ξ_1, η_1, ζ_1 when x passes through a maximum, and Equation [57] is replaced by

$$\xi=\frac{\xi_1}{\lambda_1}+\frac{\eta_1}{\lambda_2}+\frac{\zeta_1}{\lambda_3}=2R\left(\frac{\xi_1}{\lambda_1}\right)+\frac{\zeta_1}{\lambda_3} \quad \ldots\ldots [68]$$

For a system of degree n with two or more complex roots, the determination of the periodic orbits can still be reduced to a solution of Equation [51] provided $\pm a_i$ denote the increments in the normal variables ζ_i in taking up the slack, so that a pair of a_i corresponding to conjugate complex λ_i are conjugate complex. Or else a real equation may be obtained by combining conjugate pairs of the normal variables.

7 Lost Motion Taken Up Gradually

A study has been made of the motion when the slack is taken up gradually by "coasting" so that in Fig. 3, bb' is replaced by a vertical line along which $p^3\theta$ (or $p\theta$ for the second-order system) stays constant. Owing to space limitations this will be omitted except for stating the results. For $n = 2$ and real λ_1, λ_2 the hunting motions disappear and all orbits tend to rest. For complex λ_1, λ_2 the hunting motion persists, though the constant vector translations of Fig. 10 are modified.

BIBLIOGRAPHY

1 "The Effect of Backlash and of Speed-Dependent Friction on the Stability of Closed-Cycle Control Systems," by A. Tustin, *Journal of The Institution of Electrical Engineers*, vol. 94, part IIA, 1947, pp. 143–151.

2 "Instrument Inaccuracies in Feed-Back Control Systems With Particular Reference to Backlash," by H. T. Marcy, Morris Yachter, and Jerome Zauderer, Trans. AIEE, vol. 68, part I, 1949, pp. 778–788.

3 "Backlash in a Velocity Lag Servomechanism," by N. B. Nichols, Trans. AIEE, vol. 72, part II, 1953, pp. 462–467.

4 "Analysis of Backlash in Feedback Control Systems With One Degree of Freedom," by L. M. Vallese, AIEE publication "Applications and Industry," March, 1955, pp. 1–4.

Nonlinear Integral Compensation of a Velocity-Lag Servomechanism With Backlash

By C. N. SHEN,[1] H. A. MILLER,[2] AND N. B. NICHOLS[3]

When backlash is encountered in a velocity-lag servomechanism, the system may become unstable. The loop-gain limitation imposed is becoming unacceptable with increasing requirements for following steep ramp inputs as in certain machine-tool control applications. Integral compensation is not applicable since, with backlash, oscillations result. This paper analyzes an attempt to overcome this difficulty by intentionally incorporating a second nonlinearity, a dead zone in the input to the integrator. Results are given which allow the approximate transient response of certain systems to be obtained quickly and the question of stability is considered. Finally, an example of its application is given along with a discussion of several practical factors involved.

Introduction

VARIOUS investigators have used linear-system compensation techniques to obtain acceptable performance from a number of nonlinear feedback systems. To date, however, only a few efforts have been reported in utilizing nonlinear compensation methods in nonlinear systems in an effort to obtain superior results. McDonald (1)[4] and others (2 to 5) have worked on various variable damping systems as well as on the so-called "dual-mode servo." Sherrard (6) has designed a nonlinear filter to stabilize an oscillating system while Markusen and Keeler (7) used a nonlinear filter to discriminate against noise in a nonlinear system. Truxal (8) discusses nonlinear damping methods with the aid of the phase plane and Chestnut and Mayer (9) utilize frequency-response methods to study variable gain and variable time-constant systems.

The problem of backlash is a common one in servomechanism practice (10). When it is encountered in a simple velocity-lag system, which normally is absolutely stable, the system may become unstable (11). An upper gain limitation is imposed. Even without backlash the allowable gain is limited by the requirement of adequate damping. In many practical situations these limitations are acceptable but with increasing requirements for following steep ramp inputs (such as is becoming common in machine-tool control) the system error resulting from the limited loop gain may be excessive. Integral compensation is applicable to such a situation when backlash is absent and reduces the steady-state actuating signal to zero. With the backlash present, however, an oscillation limit cycle results for any positive (db) loop gain. This paper analyzes an attempt to overcome this difficulty by intentionally incorporating a second nonlinearity, a dead zone in the input to the integrator. The resulting piecewise linear set of differential equations is studied and the numerical results are given in a form which allows the approximate transient response of the system to be conveniently obtained. The question of stability is considered and, finally, an example of the application of this compensation method is given along with a discussion of several practical factors involved.

[1] Assistant Professor of Mechanical Engineering, Thayer School of Engineering, Dartmouth College, Hanover, N. H.

[2] Research Staff Member, Raytheon Manufacturing Company, Waltham, Mass.

[3] Manager, Commercial Engineering Department, Raytheon Manufacturing Company, Waltham, Mass. Mem. ASME.

[4] Numbers in parentheses refer to the Bibliography at the end of the paper.

Contributed by Instruments and Regulators Division of THE AMERICAN SOCIETY OF MECHANICAL ENGINEERS and presented at the ASME-AIEE Conference on Nonlinear Control Systems, Princeton, N. J., March 26-28, 1956.

NOTE: Statements and opinions advanced in papers are to be understood as individual expressions of their authors and not those of the Society. Manuscript received at ASME Headquarters, January 3, 1956. Paper No. 56—IRD-3.

The Nonlinear System and Equations

The nonlinear system to be considered is shown in Fig. 1 and is the mathematical model of what is considered a significant problem. It is a piecewise linear system with no saturation, coulomb friction, or noise effects. Loading and coasting effects at θ_o are neglected. It is a velocity-lag servomechanism with backlash and a proportional plus integral amplifier which has a dead zone of range $+E_d$ to $-E_d$ in series with the integrator input. The integrator is assumed ideal; i.e., with no input the output remains constant. A more practical nonideal integrator, such as a resistance-capacitance network, would introduce another parameter and is not considered quantitatively in this paper.

The motor angle θ_m for a constant-field armature-controlled electric motor is related to the manipulated voltage E_m by the following differential equation

$$T_m \frac{d^2\theta_m}{dt^2} + \frac{d\theta_m}{dt} = K_m E_m \quad \ldots\ldots [1]$$

E_m is a function of the actuating signal and depends on the action of the nonlinear compensating circuit. The integrator input is given by

$$E_1 = \begin{cases} E_{1_0} & |E| < E_d \\ \dfrac{1}{T}\displaystyle\int_0^t (E - E_d)dt + E_{1_0} & E > E_d \\ \dfrac{1}{T}\displaystyle\int_0^t (E + E_d)dt + E_{1_0} & E < -E_d \end{cases} \quad \ldots\ldots [2]$$

where T is the integral time of the integrator, E_{1_0} the initial value of the integrator output, as indicated by the second subscript, and E_d a positive constant which denotes the width of the dead zone. Then

$$E_m = E + E_1 \quad \ldots\ldots [3]$$

The relation between E and the actuating signal θ_e is defined as

$$E = K_e \theta_e \quad \ldots\ldots [4]$$

where K_e is a positive constant.

The effect of backlash is given by Equation [5] and shown graphically in Fig. 2. The letter δ represents half the free play.

As shown, it is assumed that there is no coasting of the output θ_c during reversals of the motor

$$\left.\begin{aligned} \theta_c &= \theta_m - \delta \frac{\frac{d\theta_c}{dt}}{\left|\frac{d\theta_c}{dt}\right|} & 0 < \left|\frac{d\theta_c}{dt}\right| \\ \frac{d\theta_c}{dt} &= 0 & |\theta_c - \theta_m| < \delta \end{aligned}\right\} \quad \ldots\ldots [5]$$

Dynamics and Linear Regions of the System

The dynamics of the system are complicated by both the backlash and the dead zone. Normally there are four possible regions in each of which a linear differential equation governs the motion. As is usual in piecewise linear systems the terminal conditions of one region become the initial conditions of the following region. These regions are shown in Fig. 3. For convenience the initial values of the motor angle and its rate of change are defined as

$$\theta_m|_{t=t_0} = \theta_{mo} = \theta_{co} + \delta \quad \ldots\ldots [6]$$

$$\left.\frac{d\theta_m}{dt}\right|_{t=t_0} = 0 \quad \ldots\ldots [7]$$

and

$$\theta_r = 0 \quad \ldots\ldots [8]$$

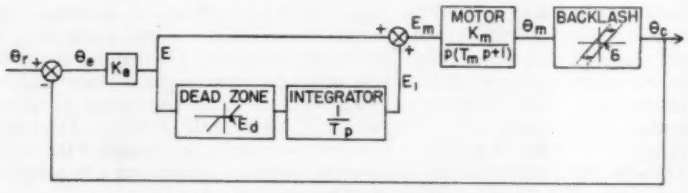


Fig. 1 The Nonlinear System With Nonlinear Integral Control

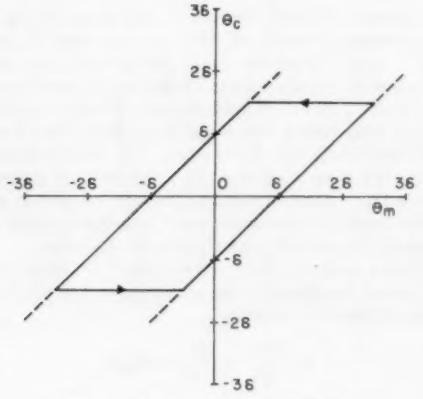


Fig. 2 Backlash Characteristic

In region A the motor angle takes up the backlash until $\theta_m = \theta_{m_0} - 2\delta$ while the output angle θ_c remains constant at θ_{co}. Then θ_m leads θ_c by the amount δ in regions B, C, and D. A third-order differential equation applies in region B until the integrator stops functioning temporarily after $E = -E_d$. Since there is no integration while $-E_d < E < +E_d$, the motion in region C is governed by a second-order differential equation and ends when $E = +E_d$. A third-order differential equation takes over in region D until the time derivative of the motor angle and hence the output angle is zero. This completes a half cycle of oscillation; a similar technique can be repeated for successive half cycles. The detailed equations are given in normalized form in the Appendix.

Complications arise since it is not known beforehand whether the system will terminate ($d\theta_m/dt = 0$) or pass through a given region. An understanding of the behavior is perhaps best obtained by showing examples for various initial conditions.

The motor and output angles, as well as the voltage E versus time are shown in Fig. 4, for a case where E initially lies outside the integrator dead zone. The backlash is taken up in region A; the system proceeds through the third-order region B and θ_m becomes a minimum in the second-order region C. Fig. 5 shows similar curves for another case where E initially lies outside the integrator dead zone, but because the initial integrator output is large θ_c becomes a minimum in region B. In Fig. 6 the initial value of E is within the integrator dead zone. The system starts

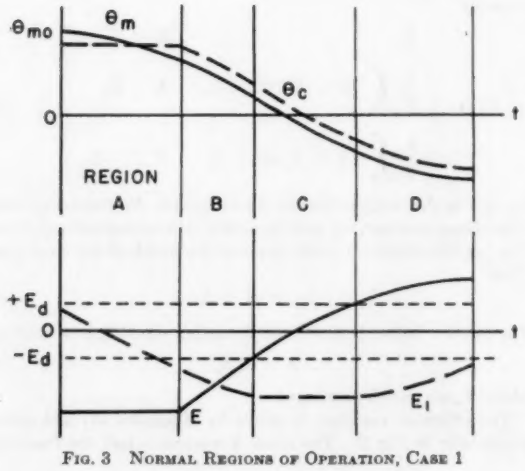


Fig. 3 Normal Regions of Operation, Case 1

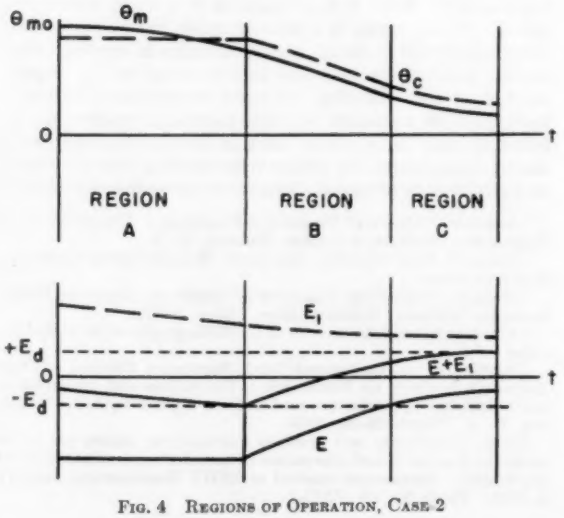


Fig. 4 Regions of Operation, Case 2

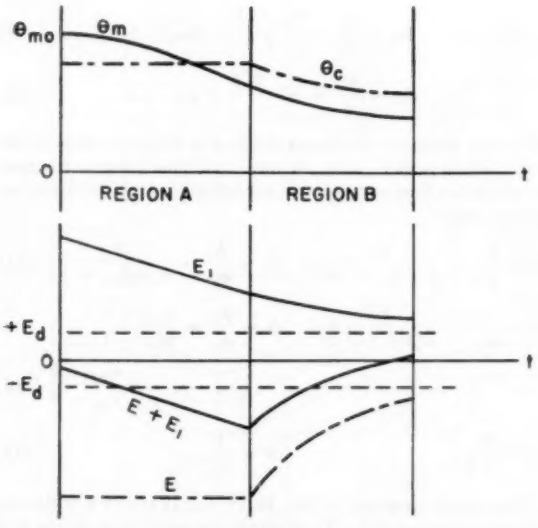


FIG. 5 REGIONS OF OPERATION, CASE 3

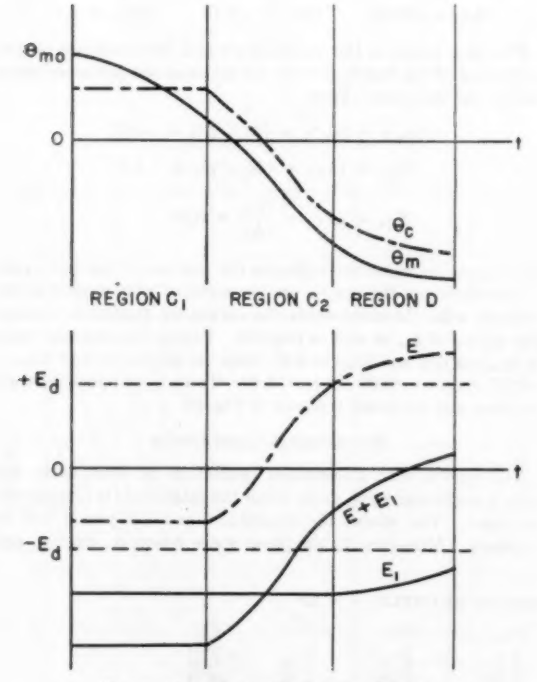


FIG. 6 REGIONS OF OPERATION, CASE 4

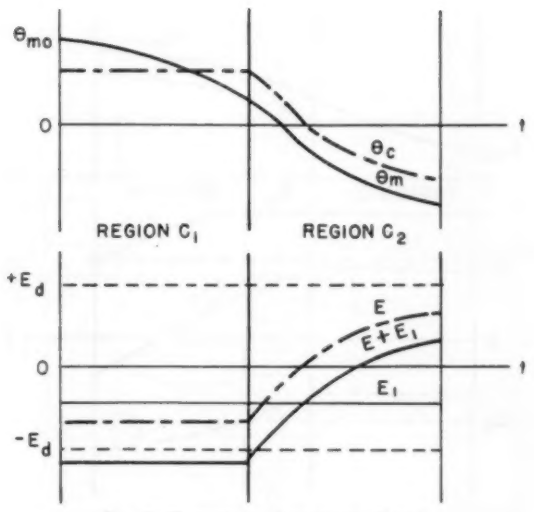


FIG. 7 REGIONS OF OPERATION, CASE 5

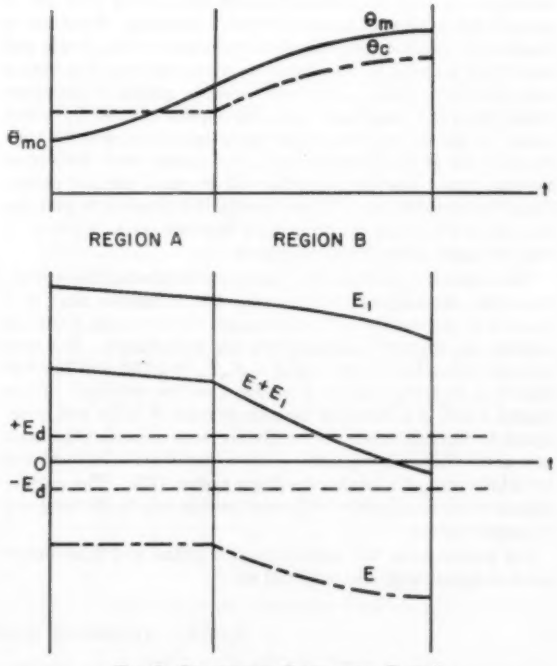


FIG. 8 REGIONS OF OPERATION, CASE 6

in region C and terminates in region D. In Fig. 7 the initial value of E is again within the integrator dead zone but the integrator output voltage is such that the system both starts and terminates in region C. This case is completely described by second-order differential equations. In Fig. 8, E again lies outside the integrator dead zone. The initial value of the integrator output is sufficiently large and of the proper sign to cause the initial acceleration to be positive. This takes up backlash in region A and terminates in the third-order region B, at a maximum output angle. The situation in Fig. 9 is similar to that in Fig. 8 except that E is within the integrator dead zone.

There are a few cases where the full amount of backlash cannot be taken up and thus a reversal of direction follows. While these cases can be analyzed on a full-cycle basis this will not be undertaken in this paper.

NUMERICAL RESULTS

To describe fully the transient solution of the system of dif-

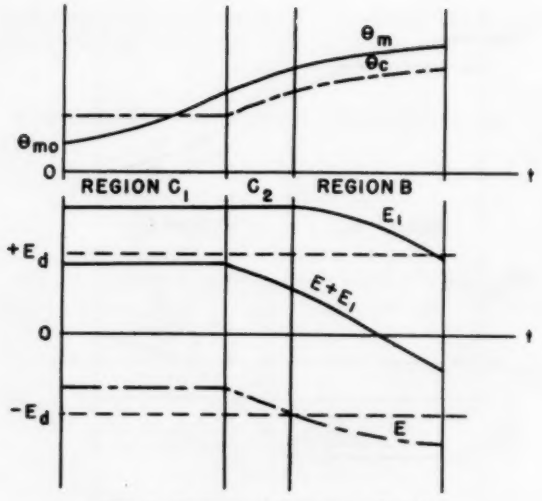


FIG. 9 REGIONS OF OPERATION, CASE 7

ferential equations requires a detailed solution for each set of parameters and for each set of initial conditions. To obtain a reasonably simple description of the behavior of this system and one which is useful in describing the transients resulting from a wide variety of initial conditions, a set of graphs is presented which gives the magnitude ratio (final value divided by initial value) of the output, the output of the integrator, and the time duration all at the moment when the output shaft derivative becomes zero. The time duration will be called the half period. Thus, by repeated use of these graphs, it is possible to plot the maxima and minima of the output function as a function of time for many given initial conditions.

The number of parameters which must be handled has been reduced by choosing the system gain and integrator time in a manner to give satisfactory performance for very large actuating signals, i.e., when the nonlinearities are insignificant. For such a linear system it is known that if $K_s K_m T_m = 1.333$, and $T_m/T = 0.2778$, a damping ratio of $\zeta = 0.316$ will be obtained. These figures result in a full-cycle magnitude ratio of 0.125 and correspond to the case where the oscillating term decays at the same rate as the nonoscillating term. This is considered to be a reasonable adjustment of a third-order linear system (12). The computations in the remainder of this paper pertain only to the foregoing numerical values.

For convenience, the symbols for the initial and final dimensionless output angle are redefined as

$$\phi_{co} = \frac{\theta_{co}}{\delta} = \frac{\theta_{mo} - \delta}{\delta} = \phi_{mo} - 1 \quad [9]$$

$$\phi_{cf} = \frac{\theta_{cf}}{\delta} = \frac{\theta_{mf} + \delta}{\delta} = \phi_{mf} + 1 \quad [10]$$

where the subscript f indicates the final or terminal value at the end of a half period. Also, the initial and final integrator output as well as the final output angle are defined in terms of the initial output angle

$$\beta_o = \frac{\lambda_o}{\phi_{co}} = \frac{E_{1o}}{K_s\delta(\phi_{mo} - 1)}, \qquad \beta_f = \frac{\lambda_f}{\phi_{cf}} = \frac{E_{1f}}{K_s\delta(\phi_{mf} + 1)} \quad [11]$$

$$\gamma = \frac{\lambda_f}{\phi_{co}} = \frac{E_{1f}}{K_s\delta(\phi_{mo} - 1)}, \qquad \Delta = \frac{\phi_{cf}}{\phi_{co}} = \frac{\phi_{mf} + 1}{\phi_{mo} - 1} = \frac{\theta_{mf} + \delta}{\theta_{mo} - \delta} \quad [12]$$

$$\eta = \frac{E_d}{K_s\delta} \qquad \sigma = \frac{t}{T_m} \quad [13]$$

The results presented in Figs. 10, 11, and 12 are for $\eta = 0.2$ and for several values of β_o. To illustrate the use of these curves it is convenient to show an example. Let $\eta = 0.2$, $\phi_{co} = 10.0$, and $\beta_o = 0$. From the graphs the following is first obtained

$$(\Delta)_1 = -0.60 \qquad (\gamma)_1 = -0.17 \qquad (\sigma_f)_1 = 3.11$$

The final values of the output angle and the integrator output at the end of the first half cycle are taken as the initial values of the second half cycle. Thus

$$(\phi_{co})_2 = (\phi_{cf})_1 = (\phi_{co})_1(\Delta)_1 = -6.0$$

$$(\lambda_o)_2 = (\lambda_f)_1 = (\phi_{co})_1(\gamma)_1 = -1.7$$

$$(\beta_o)_2 = (\beta_f)_1 = \frac{(\gamma)_1}{(\Delta)_1} = 0.28$$

The numerical subscript indicates the number of the half cycle.

The curves on the graphs are symmetrical with respect to the ordinate axis. In other words, the curves are applicable to negative values of ϕ_{co} as well as positive. Taking the absolute value of $(\phi_{co})_1$ as 6.0, and $(\beta_o)_2$ as 0.28, from the graphs we find $(\Delta)_2 = -0.37$, $(\gamma)_2 = -0.03$, $(\sigma_f)_2 = 3.40$. Table 1 continues the computation and the result is shown in Fig. 13.

SYMMETRICAL LIMIT CYCLE

The system with a sustained oscillation or limit cycle will have a particular wave shape when the magnitude is plotted versus time. The shape, the magnitude, and the period will be constant. Moreover, in the phase space where ϕ_c, $d\phi_c/d\sigma$, and

TABLE 1 TRANSIENT COMPUTATION EXAMPLE; $\eta = 0.2$

n	$(\Delta)_{n-1}$	$(\gamma)_{n-1}$	$(\phi_{co})_n$	$(\lambda_o)_n$	$(\beta_o)_n$	$(\sigma_f)_n$	$\Sigma\sigma_f$
1	...	...	10.0	0	0	3.11	3.11
2	−0.60	−0.17	−6.0	−1.7	0.28	3.40	6.51
3	−0.37	−0.03	2.2	0.18	0.081	3.63	10.14
4	−0.66	−0.17	−1.5	−0.37	0.26	4.13	14.27
5	−0.66	−0.16	0.99	0.24	0.24	4.60	18.87
6	−0.80	−0.23	−0.79	−0.23	0.29	5.14	24.01
7	−0.84	−0.25	0.66	0.20	0.30	5.50	29.51
8	−0.87	−0.28	−0.57	−0.18	0.32	5.92	35.43
9	−0.88	−0.28	0.50	0.16	0.32	6.28	41.71
10	−0.89	−0.30	−0.45	−0.15	0.34	6.72	48.43
11	−0.88	−0.29	0.40	0.13	0.33	7.16	55.59
12	−0.87	−0.29	−0.35	−0.12	0.33	7.83	63.42
13	−0.84	−0.27	0.29	0.095	0.32	8.90	72.32
14	−0.74	−0.20	−0.21	−0.058	0.27	12.40	84.72
15	−0.20	0.14	0.042	−0.029	−0.70	...	...
16	−1.5	−0.70	−0.063	−0.029	0.47	...	...
17	0.24	0.47	−0.015	−0.030	2.0	...	...
18	2.44	2.0	−0.037	−0.030	0.82	...	...
19	0.742	0.82	−0.027	−0.030	1.11	...	...

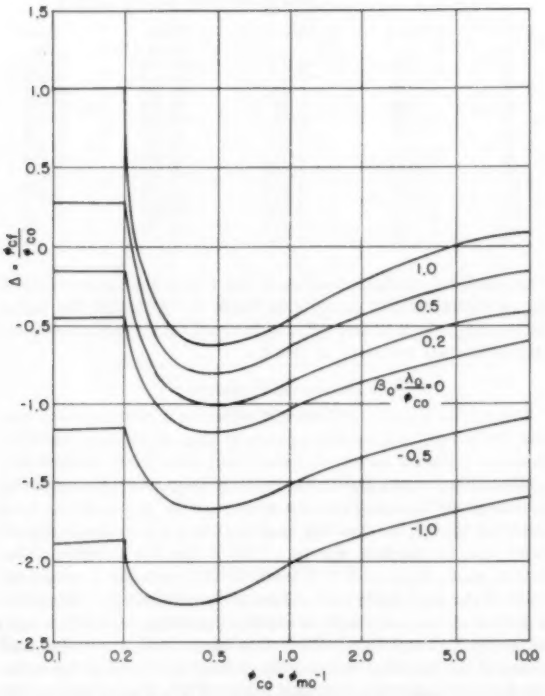


FIG. 10 MAGNITUDE RATIO FOR $\eta = 0.2$

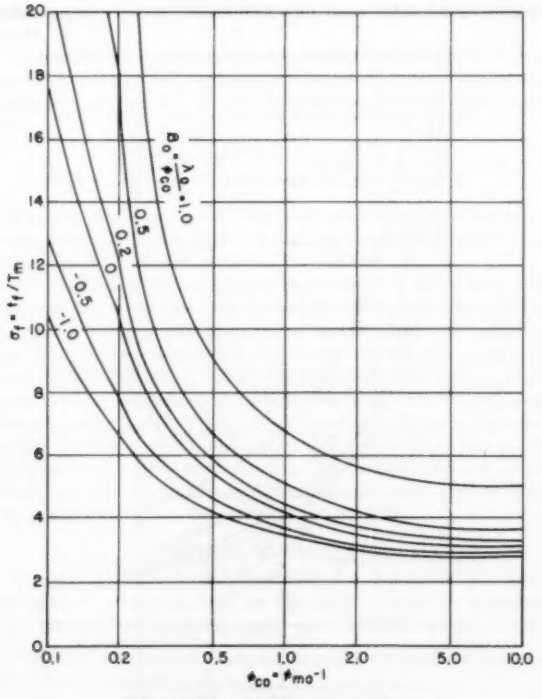


FIG. 12 HALF PERIOD FOR $\eta = 0.2$

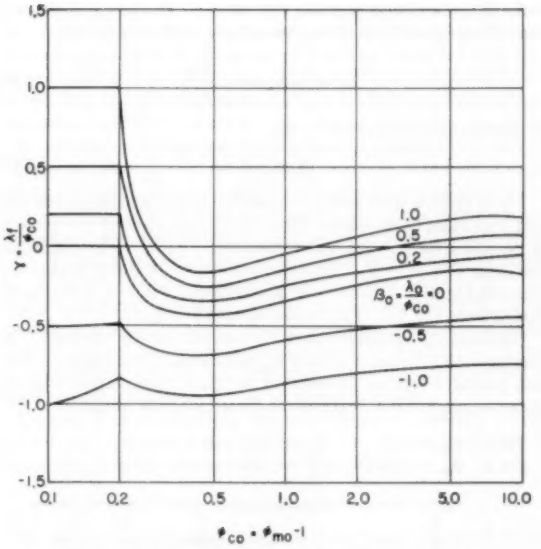


FIG. 11 INTEGRATOR OUTPUT RATIO FOR $\eta = 0.2$

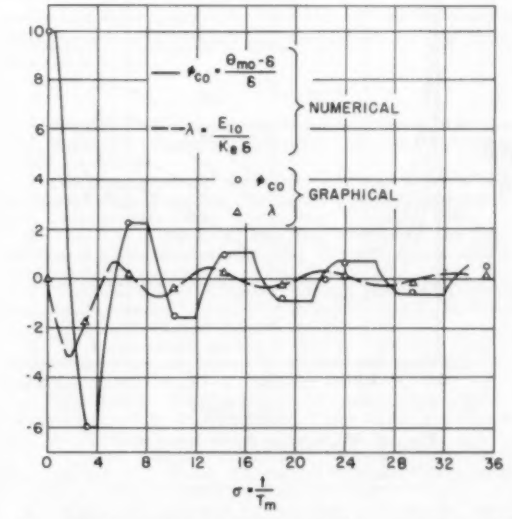


FIG. 13 COMPARISON OF TRANSIENT RESULTS FOR $\eta = 0.2$

$d^2\phi_e/d\sigma^2$ are variables, the limit-cycle curve is symmetrical with respect to the origin in this space. For a symmetrical system the magnitude ratio after a half-cycle in ϕ will be —1, i.e.

$$\phi_{ef} = -\phi_{eo} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots [14]$$

or

$$\Delta = \frac{\phi_{ef}}{\phi_{eo}} = -1$$

Since the shape of the curve remains the same the acceleration also remains the same after a full period, and after a half cycle the

ratio of the accelerations also will be —1. It also can be shown that the output of the integrator after a half cycle is

$$\lambda_f = -\lambda_o = -\frac{E_{1o}}{K_v \delta} \quad \ldots\ldots [15]$$

If this quantity is divided by ϕ_{co}

$$\gamma = -\beta_o \quad \ldots\ldots [16]$$

For a system with sustained oscillation the two Relations [14] and [16] must be satisfied. Figs. 14, 15, and 16 show the plots of Δ, γ, and σ for $\eta = 0.1$ and various β_o. In Fig. 15 we can locate the locus of $\gamma = -\beta_o$. This curve is shown on the β_o versus ϕ_{co}-plane in Fig. 17. The curve of β_o versus ϕ_{co} at $\Delta = -1$ can be obtained from Fig. 14. This is also plotted in Fig. 17.

If sustained oscillation exists, the two curves will intersect. The corresponding values of β_o and ϕ_{co} can be read immediately. The half period σ can be determined by these values from Fig. 16. As an example for $\eta = 0.1$ two intersections are obtained.

Set 1: $\phi_{co} = 0.71$ $\beta_o = 0.31$ and $\sigma = 5.1$
Set 2: $\phi_{co} = 0.128$ $\beta_o = 0.36$ and $\sigma = 14.4$

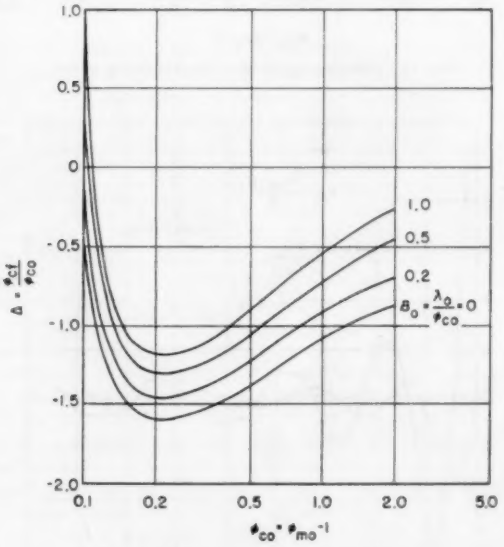


FIG. 14 MAGNITUDE RATIO FOR $\eta = 0.1$

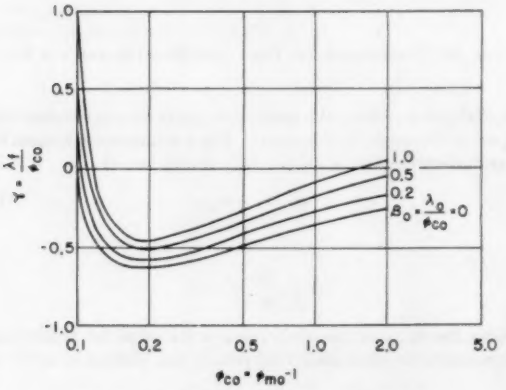


FIG. 15 INTEGRATOR OUTPUT FOR $\eta = 0.1$

TABLE 2 CONVERGENCE OF THE OSCILLATION

n	$(\Delta)_{n-1}$	$(\gamma)_{n-1}$	$(\phi_{co})_n$	$(\beta_o)_n$
0	...	...	0.13	0.30
1	−1.08	−0.42	−0.1405	0.389
2	−1.16	−0.44	0.163	0.379
3	−1.30	−0.51	−0.212	0.393
4	−1.36	−0.54	0.288	0.397
5	−1.31	−0.49	−0.377	0.374
6	−1.24	−0.44	0.467	0.355
7	−1.16	−0.38	−0.542	0.328
8	−1.09	−0.355	0.591	0.326
9	−1.07	−0.335	−0.633	0.313
10	−1.04	−0.320	0.659	0.308
11	−1.03	−0.315	−0.679	0.306
12	−1.02	−0.305	0.692	0.300
13	−1.02	−0.310	−0.705	0.304
14	−1.01	−0.305	0.712	0.302

The sustained oscillation occurs at Set 1 with the higher value of ϕ_{co} as shown by the example in Table 2. Although the initial values, $(\phi_{co})_0 = 0.13$ and $(\beta_o)_0 = 0.30$ are very close to Set 2 yet oscillations are sustained at Set 1.

RELATIVE STABILITY

One of the important factors in selecting a compensation system for a feedback control system is that of relative stability. It is not sufficient merely to assure that the system is stable in a mathematical sense, but it also must possess adequate damping. In examining transients of the system under study it has been observed that β_o is relatively constant for a few cycles in almost every case for medium signals. This is true for a region of the output shaft angle and it is possible to determine a maximum value of the magnitude ratio Δ for a truly constant β_o. When this is done it seems reasonable to use this maximum magnitude ratio as an index of stability. With this assumption, successive half cycles of the transient decrease by at least the value of the maximum magnitude ratio each half cycle. Thus, if this ratio is less than unity, the system is stable, and if it is much less than unity it is adequately damped. When this calculation is carried out for various values of the dead zone η it is seen that for adequate sta-

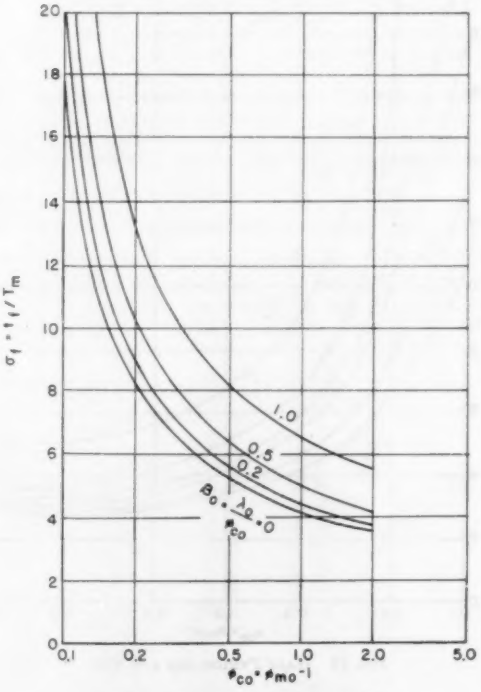


FIG. 16 HALF PERIOD FOR $\eta = 0.1$

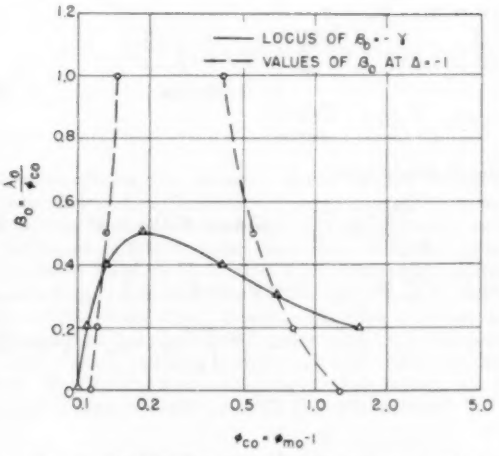


FIG. 17 DETERMINATION OF LIMIT CYCLE FOR $\eta = 0.1$

bility in a great many systems, with the gain parameters chosen in this paper, that η must be greater than 0.5. For $\eta = 0.5$, the maximum magnitude ratio is 0.58.

THE OFFSET PROBLEM

Careful examination of Table 1 will reveal that as time increases the integrator output finally achieves a constant value. This is shown in Table 1 in the column labeled λ_o and the terminal value is —0.30. This is called offset and will correspond to a final output shaft angle different from zero.[8] The magnitude of the offset is the value of the integrator output when $E/K_s\delta$ enters the region between $+\eta$ and $-\eta$ without subsequently going out of this region. With an ideal integrator the output remains at this value.

If the offset is greater than η eventually it will cause $E/K_s\delta$ to exceed η and this will reduce the integrator output. Hence the maximum offset is η. Also, if the initial conditions of the system are

$$|\phi_{mo} - 1| < \eta, \qquad \left(\frac{d\phi}{d\sigma}\right)_{\sigma=\sigma_o} = 0, \qquad \text{and } \beta_o = 1.0$$

the system will be at rest and the backlash will not be taken up in finite time. This is equivalent to saying that the final offset magnitude can be any value between $+\eta$ and $-\eta$.

Unfortunately, a system with a high value of η will have a wide range of offset. In this regard alone, η will be chosen as small as possible. Another possible solution to the offset problem is to use an integrator which will not retain its output indefinitely without an input signal. Thus, if a resistance-capacitance circuit is used as an integrator, the offset eventually will become zero. Presumably a relatively long decay time would be used, but even in this circumstance it is not certain that the results of the analysis presented in this paper apply. The analysis should be reexamined with this alternative integrator in mind to determine the possible effects on the results of the analysis.

EXAMPLE OF APPLICATION

In applying this compensation method to a practical situation, the system is first adjusted for the most satisfactory operation without the nonlinear compensator since for small signals the complete system operates in this mode. Thus, for a given T_m, with the backlash reduced to as small a value as practicable, K_sK_m is set. K_sK_m is the velocity error constant and, as shown by Nichols (11), the product $K_sK_mT_m$ must be less than 3.046 for stability. On the basis of the previous assumptions, T is already determined by the large signal performance and η is set for acceptable damping in the medium signal area.

In considering a single-motion machine-tool contouring control, experience indicates that the backlash in almost all machine tools can be reduced to 0.015 in. and for this situation a velocity-error constant of 20 sec^{-1} is a typical maximum. Such machines are currently desired to follow the equivalent of a ramp input of 50 ipm. Thus the steady-state error would be $50/(60 \times 20) = 0.042$ in. With the nonlinear compensation adjusted for $\eta = 0.5$, the steady-state error would be reduced to at most 0.0075 in. or 18 per cent of the value without the compensation. The steady-state error for zero slope inputs is scattered within ± 0.0075 in. if the ideal integrator is used, but with a modified integrator it seems likely that this impairment can be reduced to zero.

Looked at in another way the new system is desirable for all ramp inputs exceeding 60 $K_s\eta\delta$ ipm, if the observed stability is acceptable. For the above case, this is 9.0 ipm.

CONCLUSIONS

In certain circumstances with relatively large values of backlash and where it is necessary to follow a ramp input, this system may prove advantageous. The steady-state following error is reduced but an offset is introduced. Means are indicated which may reduce this impairment. With this method the steady-state system error is no more than half the backlash for step and ramp reference inputs.

[8] This is a result which would not be predicted by describing function analysis methods (13).

BIBLIOGRAPHY

1 "Nonlinear Techniques for Improving Servo Performance," by D. McDonald, Proceedings of National Electronic Conference, vol 6, 1950, pp. 400–421.

2 "A Topological and Analogue Computer Study of Certain Servomechanisms Employing Nonlinear Electronic Components," by R. C. Lathrop, PhD thesis, University of Wisconsin, 1951.

3 "The Use of Nonlinear Feedback to Improve the Transient Response of a Servomechanism," by J. B. Lewis, Trans. AIEE, vol. 71, part 2, 1952, pp. 449–453.

4 "Optimization of Nonlinear Control Systems by Means of Nonlinear Feedbacks," by R S. Nieswander and R. H. MacNiel, Trans. AIEE, vol. 72, part 2, 1953, pp. 262–272.

5 "A Differential-Analyzer Study of Certain Nonlinearly Damped Servomechanisms," by R. R. Caldwell and V. C. Rideout, Trans. AIEE, vol. 72, part 2, 1953, pp. 166–170.

6 "Stabilization of a Servomechanism Subject to a Large Amplitude Oscillation," by E S Sherrard, Trans. AIEE, vol. 71, part 2, 1952, pp. 312–324.

7 "A Noise Adaptive Flight Path Control System," by D. L. Markusen and R. J. Keeler, The Second Feedback Control Systems Conference, AIEE, No. S-63, 1954, pp. 115–122.

8 "Automatic Feedback Control System Synthesis," by J. G. Truxal, McGraw-Hill Book Company, Inc., New York, N. Y., first edition, 1955, pp. 653–663.

9 "Servomechanisms and Regulating System Design," by H. Chestnut and R. W. Mayer, John Wiley & Sons, Inc., New York, N. Y., vol. 2, first edition, 1955, pp. 345–364.

10 "The Effect of Backlash and of Speed-Dependent Friction on the Stability of Closed Cycle Control Systems," by A. Tustin, *Journal of the Institution of Electrical Engineers*, vol. 94, part 2A, 1947, pp. 143–151.

11 "Backlash in a Velocity-Lag Servomechanism," by N. B. Nichols, Trans. AIEE, vol. 72, part 2, 1953, pp. 462 and 466.

12 "Theory of Servomechanisms," by H. M. James, N. B. Nichols, and R. S. Phillips, McGraw-Hill Book Company, Inc., New York, N. Y., first edition, 1947, pp. 154–157.

13 "Recent Advances in Nonlinear Servo Theory," by J. M. Loeb, Trans. ASME. vol. 76, 1954, pp. 1281–1289 (see author's closure).

Appendix

The differential equations for each region with initial and final conditions are given for a normal half cycle that includes four regions, Fig. 3. The variables have been nondimensionalized.

Denote

$$\left.\begin{aligned} \sigma &= \frac{t}{T_m} & \sigma_o &= \frac{t_o}{T_m} \\ \phi_m &= \frac{\theta_m}{\delta} & \phi_{mo} &= \frac{\theta_{mo}}{\delta} \\ \lambda &= \frac{E_1}{K_s\delta} & \lambda_o &= \frac{E_{1o}}{K_s\delta} \\ \eta &= \frac{E_b}{K_s\delta} & K_b &= K_s K_m T_m \end{aligned}\right\} \quad \ldots\ldots [17]$$

Region A, Second Order

$$\frac{d^2\phi_m}{d\sigma^2} + \frac{d\phi_m}{d\sigma} = K_b\{\lambda_o - (\phi_{mo} - 1) + \frac{T_m}{T}(\sigma - \sigma_o)[\eta - (\phi_{mo} - 1)]\} \ldots [18]$$

At $\sigma = \sigma_o$, $\phi_m = \phi_{mo}$

$$\frac{d\phi_m}{d\sigma} = 0 \ldots\ldots [19]$$

At $\sigma = \sigma_1$, $\phi_m = \phi_{mo} - 2$

$$\left.\begin{aligned} &\frac{d\phi_m}{d\sigma} \\ &\frac{d^2\phi_m}{d\sigma^2} \end{aligned}\right\} \text{ continuous} \ldots\ldots [20]$$

Region B, Third Order

$$\frac{d^3\phi_m}{d\sigma^3} + \frac{d^2\phi_m}{d\sigma^2} + K_b\frac{d\phi_m}{d\sigma} + K_b\frac{T_m}{T}\phi_m = K_b\frac{T_m}{T}(\eta - 1) \ldots [21]$$

At $\sigma = \sigma_1$, $\phi_m = \phi_{mo} - 2$

$$\left.\begin{aligned} &\frac{d\phi_m}{d\sigma} \\ &\frac{d^2\phi_m}{d\sigma^2} \end{aligned}\right\} \text{ continuous} \ldots\ldots [22]$$

At $\sigma = \sigma_2$, $\phi_m = \eta - 1$

$$\left.\begin{aligned} &\frac{d\phi_m}{d\sigma} \\ &\frac{d^2\phi_m}{d\sigma^2} \end{aligned}\right\} \text{ continuous} \ldots\ldots [23]$$

Region C, Second Order

$$\frac{d^2\phi_m}{d\sigma^2} + \frac{d\phi_m}{d\sigma} + K_b\phi_m = K_b(\lambda_2 - 1) \ldots\ldots [24]$$

$$\lambda_2 = \lambda_o + \frac{T_m}{T}(\sigma_1 - \sigma_o)[\eta - \phi_{mo} + 1] + \frac{T_m}{T}\int_{\sigma_1}^{\sigma_1}(\eta - \phi_m - 1)\,d\sigma \ldots [25]$$

At $\sigma = \sigma_2$, $\phi_m = \eta - 1$

$$\left.\begin{aligned} &\frac{d\phi_m}{d\sigma} \\ &\frac{d^2\phi_m}{d\sigma^2} \end{aligned}\right\} \text{ continuous} \ldots\ldots [26]$$

At $\sigma = \sigma_3$, $\phi_m = -\eta - 1$

$$\left.\begin{aligned} &\frac{d\phi_m}{d\sigma} \\ &\frac{d^2\phi_m}{d\sigma^2} \end{aligned}\right\} \text{ continuous} \ldots\ldots [27]$$

Region D, Third Order

$$\frac{d^3\phi_m}{d\sigma^3} + \frac{d^2\phi_m}{d\sigma^2} + K_b\frac{d\phi_m}{d\sigma} + K_b\frac{T_m}{T}\phi_m = K_b\frac{T_m}{T}(-\eta - 1) \ldots [28]$$

At $\sigma = \sigma_3$, $\phi_m = -\eta - 1$

$$\left.\begin{aligned} &\frac{d\phi_m}{d\sigma} \\ &\frac{d^2\phi_m}{d\sigma^2} \end{aligned}\right\} \text{ continuous} \ldots\ldots [29]$$

At $\sigma = \sigma_f$, $d\phi_m/d\sigma = 0$ ..... [30]

If $d\phi_m/d\sigma = 0$ at any time prior to entering region D, the problem is not "normal" and the system behavior is described in the section on System Dynamics.

Flow Through Annular Orifices

By K. J. BELL[1] and O. P. BERGELIN[2]

Coefficients for the annular orifice formed between a circular disk and a cylindrical tube are reported for twenty-one orifices having disk diameter to tube diameter ratios in the range of 0.95 to 0.996, and orifice length-to-width ratios from 0.118 to 33.3. The orifice Reynolds-number range is from 2.0 to 20,000 for both tangent and concentric orientations of the disk. Comparison with previous data indicates that the results also apply to the annular orifice formed when a rod extends through a circular hole in a plate. Theoretical and semi-empirical equations are developed to predict coefficients for annular orifices.

Nomenclature

The following nomenclature is used in the paper:

B = orifice geometry constant, $B = 1 + \frac{3}{4} Z$
C = over-all annular orifice coefficient, Equation [1]
C' = point value of the annular-orifice coefficient at a given angle θ in a tangent orifice
C_c = coefficient of stream contraction in an orifice, Equation [10]
C_e = orifice coefficient for an eccentric annular orifice
D = inside diameter of shell or outside diameter of an annular orifice, ft
D_h = hydraulic diameter of an annular orifice, ft, $D_h = D - d$
F = fraction of maximum pressure recovery due to stream expansion from the *vena contracta* to full annulus area actually recovered in a given orifice
G = mass flow rate through an orifice, lb/(hr)(sq ft)
K = number of velocity heads pressure drop due to kinetic energy effects in the viscous-flow regime, Fig. 9
L = disk thickness, hence, orifice length, ft
P = pressure, psf
ΔP_f = pressure drop due to friction in the annulus, psf
ΔP_t = total pressure drop across an annular orifice, psf
R = outer radius of annular orifice, ft
Re = annular orifice Reynolds number, Equation [2]
Re′ = point value of Re at angle θ in a tangent annular orifice, Equation [31]
S = cross-isectional area of an annular orifice, sq ft
S_{vc} = cross- sectional area of stream at *vena contracta*, sq ft
V = veloc ty through annulus, fph
W = weight rate of flow, lb/hr
Z = concentric annulus length-to-width ratio, Equation [3]
Z' = point value of Z at angle θ in a tangent annular orifice, Equation [24]
a = clearance between disk and shell at a given angle θ in a tangent annular orifice, ft, Equation [22]
d = outside diameter of annular-orifice disk or inside diameter of annular orifice, ft
f_p = friction factor for flow between parallel plates of infinite width

$$f_p = \frac{\rho \Delta P_f g_c}{ZG^2}$$

g_c = conversion constant 4.17×10^8 (ft)(lb-mass)/(hr²)(lb-force)
r = inner radius of an annular orifice, ft
s_1 = distance from center of disk to a given point on inner perimeter of a tangent annular orifice, ft, Equation [18a]
s_2 = distance from center of disk to a given point on outer perimeter of a tangent annular orifice, ft, Equation [19]
ϵ = eccentricity of an annular orifice. ϵ is equal to the distance between centers of shell and disk divided by difference in radii of shell and disk. Therefore, $\epsilon = 0$ for a concentric orifice and $\epsilon = 1$ for a tangent orifice
θ = angle of displacement of a given point on periphery of a tangent orifice from point of greatest clearance, radians
μ = fluid absolute viscosity, lb/(ft)(hr)
π = numerical constant, 3.14159
ρ = fluid density, pcf
Σ = mathematical symbol of summation

Introduction

In the design of baffled shell-and-tube heat exchangers, mechanical considerations require that there be clearances between the shell and the baffles and between the tubes and the baffles. The annular orifices thus formed permit a part of the shell-side fluid to leak past the baffle, decreasing both the shell-side pressure drop and the heat-transfer coefficient. Therefore, it is desirable to know the annular orifice coefficient as a function of the orifice geometry and Reynolds number in order to allow for internal leakage during exchanger design. These coefficients are also applicable wherever annular clearances are encountered in other types of equipment; i.e., for housings around shafts and for labyrinth seals. As part of the Cooperative Research Program on Heat Exchangers at the University of Delaware, an experimental and analytical investigation of annular orifice coefficients was made. The results of this study are summarized here.

Experimental Equipment and Procedure

The experimental apparatus consisted of a cylindrical shell, several disks of various thicknesses and diameters, disk-support rods, an end plate, an approach header, manometers, thermocouples, and two separate fluid pumping and metering systems for the oil and the water. Figs. 1, 2, 3, and 4 show the orifice construction and the equipment layout. In operation the fluid was pumped through the orifice and the flow rate and pressure drop across the orifice were measured after both flow and thermal equilibrium were attained. The results expressed in terms of orifice coefficient, orifice Reynolds number, and orifice length-to-width ratio are shown in Figs. 5 through 8. The equations de-

[1] Hanford Atomic Products Operation, General Electric Company, Richland, Wash.; formerly, Research Fellow in Chemical Engineering, Department of Chemical Engineering, University of Delaware, Newark, Del.

[2] Professor of Chemical Engineering, University of Delaware, Newark, Del. Mem. ASME.

Contributed by the Heat Transfer Division and presented at the Spring Meeting, Portland, Ore., March 18–21, 1956, of The American Society of Mechanical Engineers.

Note: Statements and opinions advanced in papers are to be understood as individual expressions of their authors and not those of the Society. Manuscript received at ASME Headquarters, August 10, 1955. Paper No. 56—S-22.

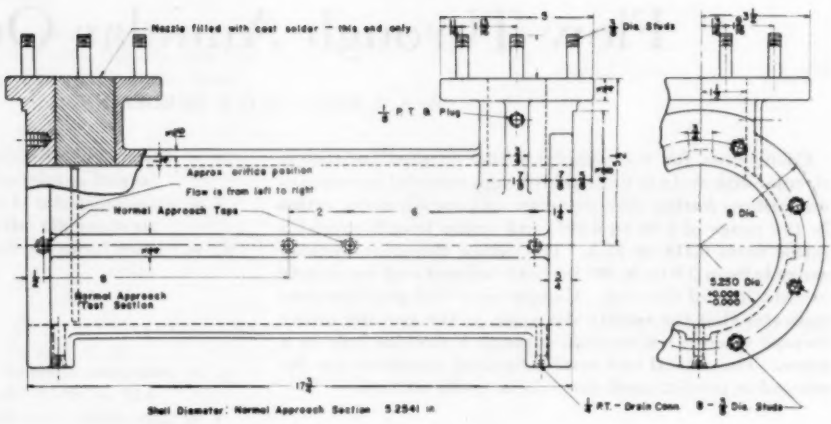
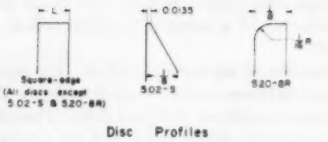


FIG. 1 DIAGRAM OF SHELL



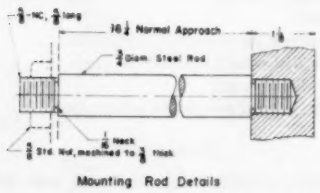


FIG. 2 MOUNTING-ROD DETAILS

fining these terms are

$$C = \frac{W}{S\sqrt{(2\rho g_c \Delta P_T)}} \quad \ldots\ldots [1]$$

$$\mathrm{Re} = \frac{(D-d)G}{\mu} \quad \ldots\ldots [2]$$

$$Z = \frac{2L}{D-d} \quad \ldots\ldots [3]$$

Details of the apparatus and procedures are presented in the Appendix. The dimensions of the orifices which were tested are given in Table 1.

THEORETICAL ANALYSIS AND EXPERIMENTAL RESULTS

In general, the orifice coefficient is a function of the orifice dimensions, the orifice Reynolds number, and the orientation of disk to shell; i.e., concentric or tangent mounting. The method of analysis which was used, was to postulate a mechanism of flow for each general case to be considered and then to separate the over-all mechanism into individual phenomena acting simultaneously. Each individual phenomenon was analyzed in the light of established principles of fluid mechanics, using rigorous methods where possible and empiricism where necessary. Finally, the results of the individual analyses were combined to predict the over-all orifice coefficient.

Space does not permit the details of the analyses (1)[3] to be shown. Hence, only the postulated mechanism and the final equation are given for each case.

In the following discussion, flow ranges are designated as viscous, turbulent, or transition, based on the Reynolds-number ranges as follows:

1 The viscous-flow range refers to Reynolds numbers less than 40, where the predominant effect is energy loss by viscous shear in the fluid, and the kinetic effects are confined to the acceleration of the fluid as it flows into the orifice.

TABLE 1 ORIFICE DIMENSIONS

	Disk diam, in.	Disk thickness, in.	Orifice area, sq ft	Z	Reynolds-number range	Type[a]
5.02-S	5.0251	0.0135	0.01284	0.118	4–22,000	C, T
5.02-16	5.0267	0.0630	0.01275	0.555	428–6560	C
5.02-16A	5.0251	0.0575	0.01284	0.502	435–6250	T
5.02-16B	5.0245	0.0650	0.01287	0.566	7100–24,500	C
5.02-16C	5.0253	0.0650	0.01283	0.569	18–21,200	C, T
5.15-16	5.1503	0.0622	0.00589	1.20	6–21000	C
5.15-16A	5.1504	0.0675	0.005885	1.30	4–21000	T
5.20-16	5.2008	0.0632	0.00304	2.38	6–18000	C, T
5.20-16A	5.2018	0.0632	0.00298	2.42	3–78	C
5.15-8	5.1486	0.1280	0.00599	2.42	6–21000	C, T
5.15-8A	5.1497	0.1280	0.00582	2.45	6–280	C
5.20-8	5.2004	0.1269	0.00306	4.73	96–3200	T
5.20-8A	5.1998	0.1269	0.00315	4.59	55–4000	C
5.20-8B	5.1998	0.1271	0.00310	4.69	7000–16,000	T
5.20-8C	5.2012	0.1271	0.00302	4.81	2–16,600	C, T
5.20-8R	5.1988	0.1269	0.00315	....	4–20,500	C, T
5.23-16	5.2309	0.0626	0.00133	5.40	2–7300	C, T
5.23-16A	5.2325	0.0626	0.00124	5.80	2–37	C
5.20-4	5.2009	0.2559	0.00303	9.61	54–5500	T
5.20-4A	5.1996	0.2559	0.00311	9.40	3–17,300	C, T
5.20-1	5.2004	0.8934	0.00447	33.3	4–17,700	C, T

[a] Tangent—T; concentric—C.
NOTE: Shell diameter was 5.2541 in. for all tests. Orifice profiles are shown in Fig. 2.

[3] Numbers in parentheses refer to the Bibliography at the end of the paper.

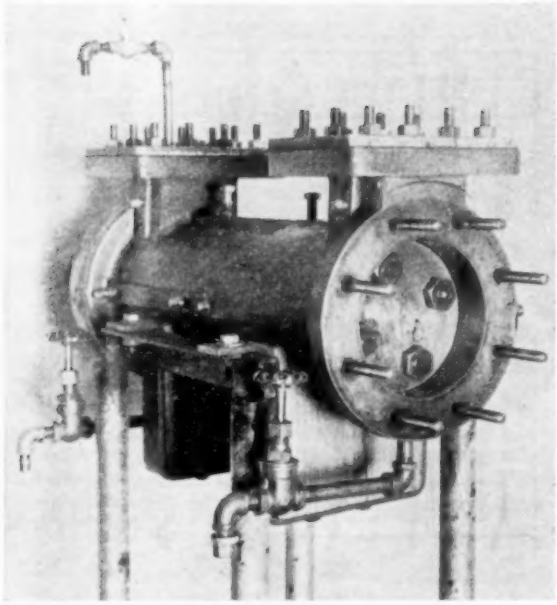


FIG. 3 TEST SECTION WITH HEADERS REMOVED

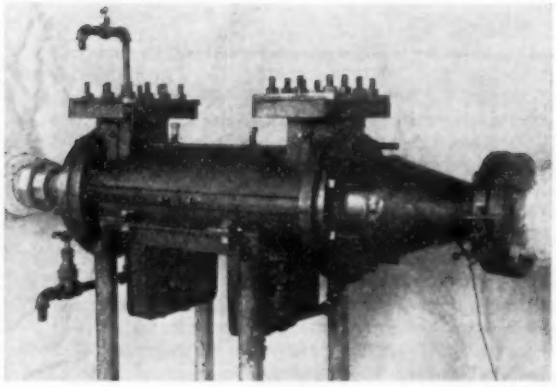


FIG. 4 APPARATUS ASSEMBLED FOR NORMAL APPROACH RUN

2 The turbulent-flow range refers to Reynolds numbers above 4000, where the predominant effects are the kinetic-energy losses associated with stream acceleration, contraction, limited expansion, and turbulent friction.

3 The transition-flow range refers to Reynolds numbers between 40 and 4000, where both kinetic and viscous phenomena are important.

There are no sharp demarcations between the flow ranges, and the foregoing division is arbitrarily based on the limits of applicability of the derived equations, as shown by the experiments.

TABLE 2 FLOW RANGES DISCUSSED

Type of orifice	Classification		
	Viscous	Turbulent	Transition
Concentric, sharp edge	A	D	
Concentric, thick, square edge	A	E	I
Concentric, thick, round edge	A	F	J
Tangent, sharp edge	B		K
Tangent, thick, square edge	C	G, H	K

Eleven separate cases, A through K, are discussed in the following sections. Table 2 serves as an index to the discussion and outlines its general scope. For each case the theoretical analysis is discussed first and, where possible, a suitable equation is developed. The experimental results are then presented and compared with the analysis. The measured coefficients for two typical orifices, 5.02-S and 5.20-1, are given in Figs. 5 and 6, and smoothed cross plots of all the data are given in Figs. 7 and 8.

(*A*) *Viscous Range, Concentric Orifice.* The mechanism of flow through an annular orifice in the viscous regime is assumed to consist of six steps:

1 Viscous deformation of the flow profile immediately upstream from the orifice.

2 Acceleration of the stream to a uniform velocity at the orifice entrance.

3 Development of a parabolic velocity distribution. The theoretical terminal-velocity distribution for viscous flow in an annulus is not exactly parabolic, but for the orifices tested, the deviation from a parabolic distribution is negligible.

4 Viscous flow with friction losses through length of orifice.

5 Dissipation of the kinetic energy of the stream downstream from the orifice.

6 Downstream viscous deformation of the stream. The equation for this case can be shown to be

$$\frac{1}{C^2} = \frac{64}{\mathrm{Re}} + \frac{48Z}{\mathrm{Re}} + K \quad [4]$$

where K is taken from Fig. 9. The first term on the right side of Equation [1] gives the effect of steps 1 and 6, the second term allows (2) for step 4, and the last term, which is negligible for $\mathrm{Re} < 4$, represents the effect of steps 2, 3, and 5. Equation [4] is also applicable to an orifice whose disk has a rounded leading edge, if the rounded length is neglected in calculating the width to length ratio Z.

For Reynolds numbers below 10, theory and experiment agree within about 2 per cent, except for a few cases in which there was considerable difficulty in centering the disk in the shell. It is believed that the theory is exact in this range and that the error was due to eccentricity of mounting. An approximate equation, the exactness of which increases with increasing Z, for the prediction of the effect of eccentricity is (4)

$$\frac{C_e}{C} = \sqrt{\left(1 + \frac{3}{2}\epsilon^2\right)} \quad [5]$$

For $10 < \mathrm{Re} < 40$, the theoretically predicted coefficients fall up to 10 per cent under the experimental results for reasons presently unknown. In this range it is recommended that the experimental results be used with due regard to possible eccentricity.

(*B*) *Viscous Range, Tangent, Sharp-Edge Orifice.* The analysis of the annular orifice formed when the disk is tangent to the shell is somewhat more complex than the concentric-orifice analysis because the clearance between disk and shell varies from $(D - d)$ to zero. Consequently, the point values of Re, Z, and C vary. The basic assumption made in the analysis of the tangent orifice is that the orifice coefficient at a given point along the periphery of a tangent orifice is the same as the coefficient of a concentric orifice having the same values of Re and Z as exist at the given point.

The problem is then one of expressing the geometric variables as functions of position along the periphery relative to the point

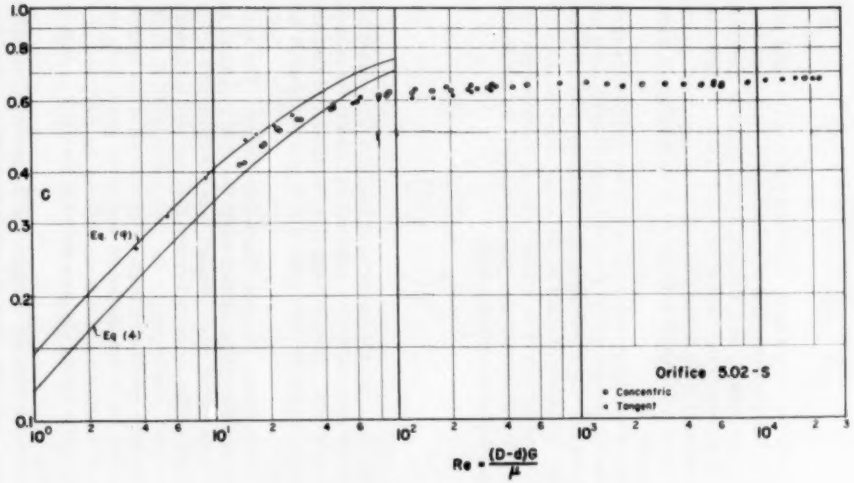


FIG. 5 COEFFICIENTS FOR ORIFICE 5.02-S

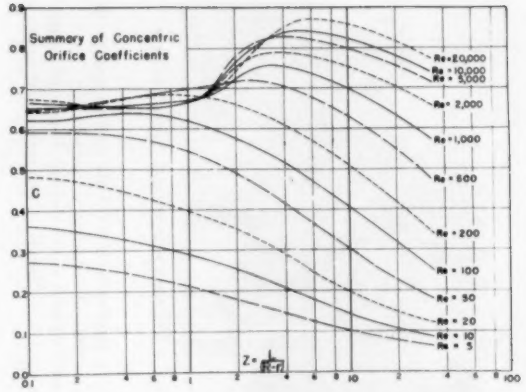


FIG. 7 SUMMARY OF CONCENTRIC-ORIFICE COEFFICIENTS

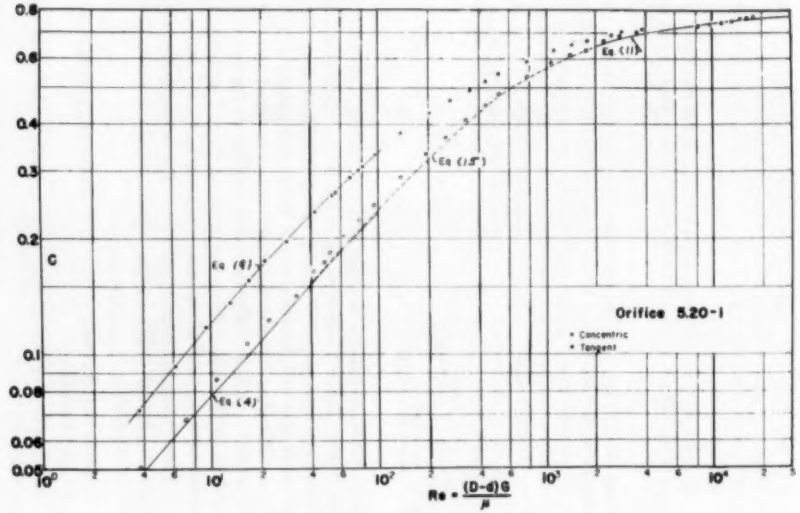


FIG. 6 COEFFICIENTS FOR ORIFICE 5.20-1

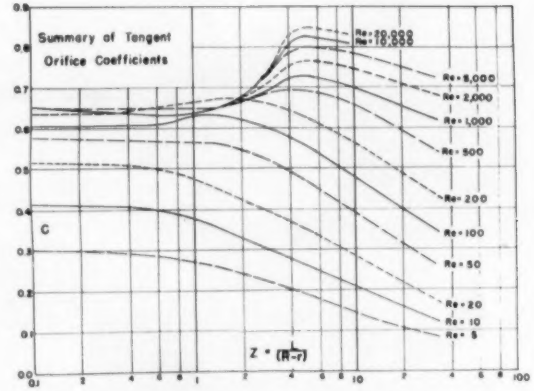


FIG. 8 SUMMARY OF TANGENT-ORIFICE COEFFICIENTS

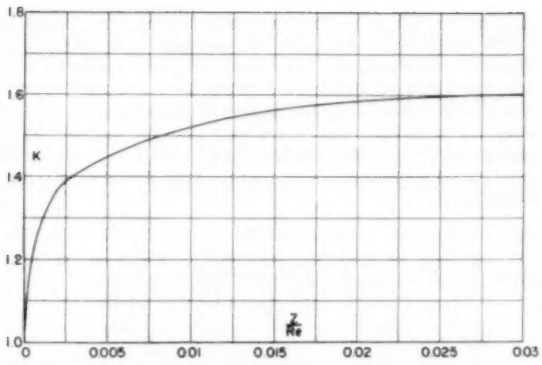


FIG. 9 KINETIC TERM FOR VISCOUS FLOW IN ANNULAR ORIFICE

of tangency, expressing the orifice coefficient as a function of these geometric variables and integrating the resulting expression over the entire orifice circumference. The detailed treatment of the case of a tangent orifice having Z equal to zero, assuming negligible kinetic effects ($K \ll 64/\mathrm{Re}$), is given in the Appendix. The resulting equation is

$$\frac{1}{C^2} = \left[\frac{64}{\mathrm{Re}}\,\frac{R+r}{2R+r}\right] \quad \ldots\ldots\ldots\ldots\ldots [6]$$

It has been found that the use of K from Fig. 9 gives fair agreement with the data for $\mathrm{Re} < 40$ for tangent orifices, and so it may be used to calculate orifice coefficients. For the case of $Z = 0$ and hence $K = 1$, the recommended equation is

$$\frac{1}{C^2} = \frac{64}{\mathrm{Re}}\left[\frac{R+r}{2R+r}\right] + 1 \quad \ldots\ldots\ldots\ldots\ldots [7]$$

or

$$\frac{1}{C^2} \doteq \frac{128}{3\mathrm{Re}} + 1$$

where the symbol $\doteq$ means "approximately."

It may be seen from Fig. 5, that Equation [7], which is a form of Equation [9], and the experimental results agree within about 1 per cent for $\mathrm{Re} < 25$. For $\mathrm{Re} > 25$, experimental data should be used.

(*C*) *Viscous Range, Tangent, Thick Orifice.* The same method of analysis as used in Case B may be used for an orifice having Z not equal to zero, again neglecting kinetic effects. The resulting expression for the orifice coefficient is

$$C^2 = \frac{\mathrm{Re}}{32(R+r)}\left[\left(\frac{5}{2}R + r\right) - B^3(R-r) + B^2(3R-2r) - \frac{B}{2}(7R-r) + \frac{BR(B-1)^3}{\sqrt{(B^2-1)}} - \frac{r(B-1)^3(B+1)}{\sqrt{(B^2-1)}}\right] \quad \ldots [8]$$

where $B = 1 + \frac{3}{4}Z$, and $Z \neq 0$

It is possible to arrive at a simpler, but not exact, expression for thick tangent-orifice coefficients by assuming that the expressions for stream-deformation losses and channel-friction losses may be independently integrated and then combined. In this form it is also convenient to include K as an empirical correction factor for kinetic effects. The result is

$$\frac{1}{C^2} = \frac{64}{\mathrm{Re}}\left(\frac{R+r}{2R+r}\right) + \frac{192}{5}\left(\frac{Z}{\mathrm{Re}}\right)\left(\frac{R+r}{3R+r}\right) + K$$

or

$$\frac{1}{C^2} \doteq \frac{128}{3\mathrm{Re}} + \frac{96}{5}\,\frac{Z}{\mathrm{Re}} + K \quad \ldots\ldots\ldots\ldots [9]$$

The maximum difference between Equation [8] and either form of Equation [9] is 0.8 per cent at $Z = 4/3$.

Equation [9] predicts experimental results within about 2 per cent for $\mathrm{Re} < 40$. As Z increases, the range of validity of Equation [9] increases. Hence, Fig. 6 shows that for $Z = 33.3$, Equation [9] is valid to $\mathrm{Re} = 100$.

(*D*) *Turbulent Range, Concentric, Sharp-Edge Orifice.* For turbulent flow through a sharp-edge orifice, the stream continues to contract and accelerate downstream until a minimum area S_{vc} is reached at the *vena contracta*. The stream then expands and decelerates, and the kinetic energy is dissipated as friction. For the orifices studied, the downstream area is very much greater than S_{vc}, the pressure recovery is negligible, and the total pressure loss is equal to that required to accelerate the fluid to the *vena contracta* plus any entrance-friction losses. Entrance-friction losses are generally considered to be not over 1 to 2 per cent of the gain in stream kinetic energy, so they are conveniently included in the acceleration pressure drop. Therefore, with small error

$$C = C_c = S_{vc}/S \quad \ldots\ldots\ldots\ldots\ldots [10]$$

The value of C_c is taken from Fig. 5, a plot of the coefficient for orifice 5.02-S. In the turbulent-flow regime, orifice 5.02-S behaves as a sharp-edge orifice even though it has an edge of finite, but small, thickness.

No quantitative theory was developed for this case. Values for C should be taken from Fig. 5.

(*E*) *Turbulent Range, Concentric, Thick, Square-Edge Orifice.* For the case of the thick orifice in the turbulent regime, two phenomena, in addition to the initial acceleration and contraction, occur and must be included in the analysis. These phenomena are:

1 The stream expands from the *vena contracta* to the full area of the annulus with a partial recovery of the kinetic energy as pressure. The expansion from the *vena contracta* begins a finite distance downstream from the orifice entrance, and no pressure recovery will be obtained in an orifice whose thickness is less than this distance. Also the pressure recovery is not instantaneous, but occurs over a distance. Hence, an orifice whose thickness is between the value at which pressure recovery starts and the value at which the recovery is essentially complete will benefit by only a fraction of the maximum possible pressure recovery. The nature of the "efficiency of pressure-recovery" function is determined from the experimental results.

2 Wall-friction losses occur during flow through the orifice length. The frictional losses are calculated from a friction factor versus Reynolds number curve for flow between parallel plates, the curve being based on available data (2, 7, 8). This method of evaluation of annular-orifice friction losses can be considered only approximate due to the disturbed entrance conditions, partial inclusion of frictional losses in the analysis of the pressure recovery, and lack of knowledge regarding relative roughness. The resulting equation is found to be (1)

$$\frac{1}{C^2} = \frac{1}{C_c^2} - \left[\frac{2}{C_c} - 2\right]F + 2f_pZ \quad \ldots\ldots\ldots [11]$$

where $F = 0$, for $Z < 1.15$ and $F = 1 - e^{-0.88(Z-1.15)}$, for $Z > 1.15$. C_c and f_p are taken from Figs. 5 and 10, respectively.

For most tests Equation [11] predicted results within 2 per cent for $\mathrm{Re} > 4000$. In a few cases, especially for $\mathrm{Re} < 10{,}000$, there were differences up to 4 per cent. Values should be taken from Fig. 6 where available, but Equation [11] appears to be entirely acceptable for extrapolation.

(*F*) *Turbulent Range, Concentric, Thick, Round-Edge Orifice.*

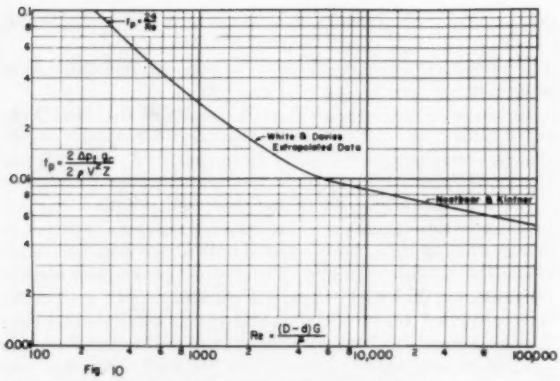


Fig. 10

FIG. 10 FRICTION FACTOR FOR ANNULI OF FINE CLEARANCE AND FOR PARALLEL PLATES

For the case of the round-edge orifice in the regime of fully developed turbulent flow, Equation [11] can be simplified considerably. If the disk edge is sufficiently rounded, there will be no stream contraction; i.e., C_c will equal unity minus any effect of entrance friction, and there will be no pressure recovery. Assuming a 1 per cent loss of kinetic energy due to entrance friction, Equation [11] becomes

$$1/C^2 = 1.01 + 2f_pZ \quad \ldots\ldots\ldots\ldots\ldots\ldots [12]$$

The rounded-entrance length is not included in the calculation of Z.

On the basis of results from one orifice only, Equation [12] agrees with experiment within 2 per cent for Re > 10,000. Equation [12] is seriously in error at lower Reynolds numbers, probably due to entrance effects which are not included in the analysis. For high flow rates and low Z-values, C can be taken as unity. At Re > 10,000, Equation [12] may be used to predict the effect of thickness.

(G) Turbulent Range, Tangent, Square-Edge Orifice. A general method of evaluating tangent-orifice coefficients is to evaluate numerically the equation

$$C = \frac{\Sigma(C'\Delta S)}{S} \quad \ldots\ldots\ldots\ldots\ldots\ldots [13]$$

from concentric experimental data. The procedure is

1 Divide the orifice into an appropriate number of sectors.

2 From Equations [27], [28], and [29] (see Appendix), determine value of ΔS, a, and Z' for each sector.

3 Assume a value C' for the sector.

4 Compute Re′ from the assumed C'.

5 Check the assumed value of C' against the experimental value from Fig. 8.

6 Assume a new C' and repeat steps 4 and 5 until agreement is attained.

7 Repeat for each sector and substitute the results into Equation [13].

Equation [13] has been evaluated for several cases, most of them for an orifice of $Z = 2.40$ where the variation of C with Re is somewhat erratic. In general, the results are within 3 per cent of experiment, and there is evidence that agreement within 2 per cent is obtained for orifices of greater Z. However, experimental results should be used when available.

(H) Turbulent Range, Tangent, Thick, Square-Edge Orifice, $Z > 9$. The integration of Equation [11] in the manner shown for Equation [6] in the Appendix for the viscous regime is attended with both theoretical and mathematical difficulties. Equation [11] applies only to fully developed turbulent flow, but in the orifice region near the point of tangency both viscous and transition regions occur. Mathematically, it is necessary to restrict the treatment to a thick orifice so that pressure recovery is effectively complete at all points around the perimeter. It is also necessary to choose mean values of f_p and C_c, evaluated at the over-all orifice Reynolds number. Within these restrictions, the equation for a tangent-orifice coefficient for $Z > 9$ and for Re > 10,000 is

$$C_T = \frac{2}{S_0}\left\{ L^2 \left\langle \left[\frac{3}{2}\left(\frac{1}{Z} - \frac{f_p}{\frac{1}{C_c^2} - \left(\frac{2}{C_c} - 2\right)} \right) \left(\frac{1}{Z\sqrt{\frac{1}{C_c^2} - \left(\frac{2}{C_c} - 2\right)}} - \frac{f_p}{\left[\frac{1}{C_c^2} - \left(\frac{2}{C_c} - 2\right)\right]^{3/2}} \right) + \left(\frac{2f_p}{Z\left[\frac{1}{C_c^2} - \left(\frac{2}{C_c} - 2\right)\right]^{3/2}} \right) \right] \left[\frac{\pi}{2} + \sin^{-1}\left(\frac{\frac{1}{C_c^2} - \left(\frac{2}{C_c} - 2\right) - Zf_p}{\frac{1}{C_c^2} - \left(\frac{2}{C_c} - 2\right) + Zf_p} \right) \right] + \left[\frac{3}{2}\left(\frac{1}{Z} - \frac{f_p}{\frac{1}{C_c^2} - \left(\frac{2}{C_c} - 2\right)} \right) \right] \left[\frac{2}{\frac{1}{C_c^2} - \left(\frac{2}{C_c} - 2\right)} \sqrt{\frac{f_p}{Z}} \right] \right\rangle + L(2r - R) \left\langle \left[\frac{1}{Z} - \frac{f_p}{\frac{1}{C_c^2} - \left(\frac{2}{C_c} - 2\right)} \right] \left[\frac{1}{\sqrt{\frac{1}{C_c^2} - \left(\frac{2}{C_c} - 2\right)}} \right] \left[\frac{\pi}{2} + \sin^{-1}\left(\frac{\frac{1}{C_c^2} - \left(\frac{2}{C_c} - 2\right) - Zf_p}{\frac{1}{C_c^2} - \left(\frac{2}{C_c} - 2\right) + Zf_p} \right) \right] + \frac{2}{\frac{1}{C_c^2} - \left(\frac{2}{C_c} - 2\right)} \sqrt{\frac{f_p}{Z}} \right\rangle \right\} \quad \ldots\ldots [14]$$

Obviously this equation is too restricted in its assumptions and too complicated to be of general use. Based on very limited data, Equation [14] appears to be applicable to thick orifices at high Reynolds numbers, but experimental results should be used where available.

(I) Transition Range, Concentric, Thick, Square-Edge Orifice. A rigorous analytical treatment of concentric orifices in the transition regime, taking into account all of the viscous and kinetic effects, is not possible at present because the exact variation of each effect with Re and Z is not known. As a first approximation, a rather rough but basically reasonable analysis has been made by

combining the elements of the treatments developed for the viscous and turbulent regimes. The resulting equation is

$$\frac{1}{C^2} = \frac{1}{C_c^2} - \left[2\sqrt{\left(\frac{1}{C_c^2} - \frac{64}{\mathrm{Re}}\right)} - 2\right]F + 2f_pZ \ldots [15]$$

$F = 0$ for $Z < 1.15$, $F = 1 - e^{-0.86(Z-1.15)}$ for $Z > 1.15$

where C_c and f_p are taken from Figs. 5 and 10 at the appropriate Reynolds number.

Equation [15] predicts coefficients within 4 per cent over most of the range studied. The greatest errors (up to 12 per cent) occur for orifices having $Z < 2.5$, when $50 < \mathrm{Re} < 400$. These errors are believed to be due to pressure recovery occurring closer to the orifice entrance than predicted by the F given in Equation [15]. In the transition range, experimental values should be used whenever possible, and Equation [15] used for extrapolation to higher values of Z.

(*J*) *Transition Range, Concentric, Thick, Round-Edge Orifice.* An approximate analysis for this case (1) gives

$$\frac{1}{C^2} = \frac{64}{\mathrm{Re}} + K + 2f_pZ \ldots [16]$$

where K and f_p are taken from Figs. 9 and 10, respectively.

Based on results from just one orifice, Equation [16] predicts a coefficient too low by 3 to 5 per cent. Equation [16] is probably satisfactory for predicting the effect of thickness and the prediction of coefficients for orifices of large Z.

(*K*) *Transition Range, Tangent, Square-Edge Orifice.* The method developed for Case G is applicable to this case, but experimental results should be used when available.

Discussion of Results

The data from all twenty-one orifices were first plotted in the manner shown in Figs. 5 and 6. These results were then summarized in the generalized plots shown in Figs. 7 and 8.

Sullivan (5) has published orifice-coefficient data for the case of a tube passing through a hole of slightly greater diameter in a thin plate, the inverse of the present case. His results agree within ±5 per cent of those reported here. Sullivan also established that coefficients for the case of flow approaching normal to the orifice, i.e., at 90 deg to the axis of the orifice, are applicable to calculation of baffle leakage in baffled heat exchangers where there are multiple orifices and the flow approaches the leakage areas from various angles. Therefore, the present results are also usable for the leakage around tubes passing through baffles.

Howell (6) studied a sharp-edged disk in a cylindrical shell with disk-to-shell-diameter ratios from 0.7 to 0.9. Extrapolation of his data to the diameter ratios studied here indicates good agreement with the present results.

In Figs. 7 and 8 the effects of pressure recovery and wall friction are clearly shown. For high Reynolds numbers the pressure recovery begins at a Z-value of about 1.0 and causes an increase in the flow rate and thus in the coefficient up to a Z-value of about 6.0. For longer orifice channels the wall friction causes a decrease in the flow rate until, at channel lengths giving Z-numbers in the range of 10 to 100, the coefficient again drops to about 0.65, the value for the sharp-edged orifice. At higher Z-values the frictional resistance lowers the value of the coefficient still more.

There is no pressure recovery for Reynolds numbers below 100 and friction causes a continual decrease in the coefficient as the orifice channel lengthens.

By the use of Figs. 7 and 8 the flow through annular openings can be predicted for a wide range of flow rates and orifice dimensions. Also, to a limited extent, these curves can serve as a design guide to enhance the flow, as might be desired for orifice-type baffled heat exchangers or for coolant around a shaft, or to minimize the flow as would be desired between the baffle and the shell of a segmentally baffled heat exchanger. The magnitude of leakage flows through the clearances generally found in commercial heat exchangers is surprisingly large, and often amounts to over half of the flow through the exchanger. With a better knowledge of the nature of these flows more accurate predictions of performance can be made and a more rational approach to design is possible.

Acknowledgment

The author acknowledges the valuable suggestions on the program of the study made by the Special Advisory Committee of the Heat Transfer Division of The American Society of Mechanical Engineers. During the course of this work, the committee included Chairman A. C. Mueller, C. H. Brooks, A. P. Colburn, S. Kopp, W. H. McAdams, B. E. Short, W. H. Thompson, T. Tinker, and P. R. Trumpler. Special thanks are due to A. Wurster for his advice on the equipment design and his comments on the design methods. The assistance of Research Fellow M. D. Leighton, who took the photographs used in this work and who aided in numerous other tasks, is greatly appreciated. Funds and equipment were furnished by Tubular Exchanger Manufacturer's Association, The American Petroleum Institute, Andale Company, and E. I. du Pont de Nemours and Company. The oils were furnished by the Gulf Oil Company. The National Science Foundation provided a fellowship under which the research was done.

BIBLIOGRAPHY

1 "Annular Orifice Coefficients With Application to Heat Exchanger Design," by K. J. Bell, PhD thesis, Department of Chemical Engineering, University of Delaware, Newark, Del., 1955.

2 "Fluid Dynamics and Heat Transfer," by J. G. Knudsen and D. L. Katz, Engineering Research Bulletin No. 37, Engineering Research Institute, University of Michigan, Ann Arbor, Mich., 1953.

3 "Laminare Kanaleinlaufströmung," by H. Schlicting, *Zeitschrift für angewandte Mathematik und Mechanik*, vol. 14, 1934, pp. 368–373.

4 "Viscous Flow Through Pipes With Cores," by N. A. V. Piercy, M. S. Hooper, and H. F. Winny, *Philosophical Magazine*, vol. 15, 1933, pp. 647–676.

5 "Heat Transfer and Fluid Friction in a Shell and Tube Exchanger with a Single Baffle," by F. W. Sullivan and O. P. Bergelin, presented at the joint AIChE-ASME Symposium on Heat Transfer at the 1955 Spring Meeting of The American Institute of Chemical Engineers.

6 "A Note on R. A. E. Annular Airflow Orifice," by A. R. Howell, Aeronautical Research Committee Reports and Memoranda No. 1934, British, 1939.

7 "An Experimental Study of the Flow of Water in Pipes of Rectangular Section," by S. J. Davies and C. N. White, Proceedings of the Royal Society of London, England, series A, vol. 119, 1928, pp. 92–107.

8 "Fluid Friction in Annuli of Small Clearance," by R. F. Nootbaar and R. C. Kintner, Proceedings of the Second Midwestern Conference on Fluid Mechanics, Ohio State University Engineering Experiment Station Bulletin No. 149, 1952, pp. 185–199.

Appendix

Derivation of Equation for a Tangent Sharp-Edge Orifice in Viscous-Flow Regime With No Kinetic Effects. The first step in the derivation is to establish the equations of the geometric variables. As shown in Fig. 11, the equations of the two circles are

$$r^2 = x^2 + y^2 \ldots [17a]$$

$$R^2 = [x - (R - r)]^2 + y^2 \ldots [17b]$$

Transforming to polar co-ordinates gives

$$r^2 = s_1^2 \ldots [18a]$$

$$R^2 = [s_2 \cos\theta - (R - r)]^2 + s_2^2 \sin^2\theta \ldots [18b]$$

where s_1 and s_2 are the distances from the origin to points on the inner and outer circles, respectively, at a given value of θ.

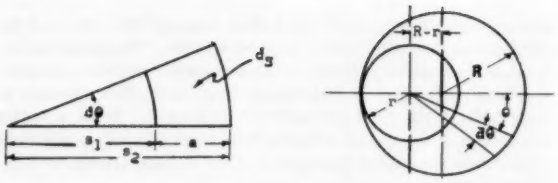


FIG. 11

Expanding, simplifying, and solving Equation [18*b*] for s_2 gives

$$s_2 = (R - r)\cos\theta + \sqrt{[(R-r)^2\cos^2\theta + 2Rr - r^2]} \quad [19]$$

The incremental area of the orifice in any sector is

$$dS = (\pi s_2^2 - \pi s_1^2)\frac{d\theta}{2\pi} \quad [20]$$

Substituting Equations [18*a*] and [19] into [20] gives

$$dS = \{(R-r)^2\cos^2\theta + (R-r)\sqrt{[(R-r)^2\cos^2\theta + 2Rr - r^2]}\cos\theta + Rr - r^2\}d\theta \quad [21]$$

The clearance between shell and disk, a, at angle θ is

$$a = s_1 - s_2 \quad [22]$$

Or using Equations [18*a*] and [19]

$$a = (R-r)\cos\theta + \sqrt{[(R-r)^2\cos^2\theta + 2Rr - r^2]} - r \quad [23]$$

The point value of Z at angle θ is given by

$$Z' = L/a \quad [24]$$

(Z', Re′, and C' will be used to denote point values of Z, Re, and C in a tangent orifice.) Combining Equations [23] and [24] gives

$$Z' = \frac{L}{(R-r)\cos\theta + \sqrt{[(R-r)^2\cos^2\theta + 2Rr - r^2]} - r} \quad [25]$$

A simplification in the equations may be made for the case where $r/R > 0.8$ by the following approximation

$$Rr \doteq \frac{R^2}{2} + \frac{r^2}{2} \quad [26]$$

Equation [21] then becomes

$$dS = [(R-r)^2\cos^2\theta + R(R-r)\cos\theta + Rr - r^2]d\theta \quad [27]$$

Equation [23] becomes

$$a = (R-r)(\cos\theta + 1) \quad [28]$$

And Equation [25] becomes

$$Z' = \frac{L}{(R-r)(\cos\theta + 1)} \quad [29]$$

In order to express the orifice coefficient in terms of the geometric variables, the incremental flow rate dW through area dS is expressed in the following form

$$dW = \sqrt{(2\rho g_c \Delta P_i)}C'\,dS \quad [30]$$

where C' and dS are functions of θ.

For the case of the concentric sharp-edge orifice ($Z = 0$), it was found that, neglecting kinetic effects

$$1/C_2 = 64/\mathrm{Re} \quad [4]$$

The point value of Re is

$$\mathrm{Re}' = \frac{2a}{\mu}\left(\frac{dW}{dS}\right) \quad [31]$$

Introducing Equations [28] and [30] into [31] gives

$$\mathrm{Re}' = \frac{2(R-r)(\cos\theta + 1)\sqrt{(2\rho g_c \Delta P_i)}}{\mu}C' \quad [32]$$

Introducing Equation [32] into [4], noting C now equals C', gives

$$C' = \frac{2(R-r)(\cos\theta + 1)\sqrt{(2\rho g_c \Delta P_i)}}{64\mu} \quad [33]$$

Substituting Equations [27] and [33] into [30] gives

$$dW = \left[\frac{4(R-r)(\rho g_c \Delta P_i)(\cos\theta + 1)}{64\mu}\right][(R-r)^2\cos^2\theta + R(R-r)\cos\theta + Rr - r^2]d\theta \quad [34]$$

Equation [34] is integrated from $\theta = 0$ to $\theta = \pi$ and the integral multiplied by 2 to allow for symmetry. Simplifying the integral yields

$$W = \frac{8\rho g_c \Delta P_i (R-r)^2}{64\mu}\int_0^\pi (1 + \cos\theta)^2[(R-r)\cos\theta + r]d\theta \quad [35]$$

Expanding the terms in the integral and integrating yields

$$W = \frac{8\rho g_c \Delta P_i (R-r)^2}{64\mu}\left[r\theta + (R-r)\sin\theta + \frac{1}{2}(2R-r)(\sin\theta\cos\theta + \theta) + \frac{1}{3}(R-r)\sin\theta(\cos^2\theta + 2)\right]\Bigg|_0^\pi \quad [36]$$

Introducing limits into Equation [36] gives

$$W = \frac{8\rho g_c \Delta P_i \pi (R-r)^2}{64\mu}\left(R + \frac{r}{2}\right) \quad [37]$$

Squaring Equation [1] and combining with Equation [37] gives

$$C^2 = \frac{4W\pi(R-r)^2}{64S^2\mu}\left(R + \frac{r}{2}\right) \quad [38]$$

where C is now the over-all coefficient for the tangent orifice. Introducing Equation [2] and $S = \pi(R^2 - r^2)$ into Equation [37] gives

$$C^2 = \frac{2\mathrm{Re}}{64}\left[\frac{R + \frac{r}{2}}{R + r}\right] \quad [39]$$

or, since R and r are very nearly the same for the orifices studied

$$C^2 = 3\mathrm{Re}/128 \quad [40]$$

Equation [7] is Equation [40] with an empirical kinetic correction term.

Apparatus. The cylindrical shell, shown in Fig. 1, was of cast bronze and was bored to an internal diameter of $5.250\ ^{+0.008}_{-0.000}$ in. The diameter at the test section was 5.2541 in. Pressure taps

were drilled through the cast pad on the side of the shell at the points indicated in Fig. 1. These points were $^3/_4$ to $^{15}/_{16}$ in. upstream, depending on disk thickness, and 9 in. downstream. The pressure-tap holes in the inside-shell wall were $^1/_{16}$ in. diam. The approximate position of the orifice plates is shown.

The disks were machined from metal plates of the desired thickness. The disk diameters ranged from 5.0245 to 5.2325 in., giving diametral clearances from 0.2296 to 0.0216 in. The disks had square edges and ranged in thickness from 0.0135 to 0.8934 in. As shown in Fig. 2, the disks were held in place by means of four $^3/_4$-in-OD steel support rods. These rods were screwed into a 1-in-thick steel end plate which in turn was held by the studs on the downstream end of the shell. When tests were to be made with concentric mounting, i.e., uniform clearance between disk and shell around the disk periphery, three support tabs were soldered to the downstream face of the disk and extended over the edge to touch the shell. When the disks were to be tested with the edge of the disk tangent to the shell, a small shim was inserted at one point between the shell and end plate to force the disk against the shell. Fig. 3 shows the shell and a disk in place with connecting piping removed, and Fig. 4 shows the apparatus ready for operation.

A conical header was used to bring the flow from the 2-in. pipe to the shell. A 40-mesh screen was inserted between the header and the shell to give a reasonably flat velocity profile.

The oil-pumping system was constructed of 2-in. steel pipe. The oil flowed from a large storage tank to one of two rotary positive-displacement pumps in parallel. The flow rate through the orifice was controlled by a by-pass line immediately after the pump. The orifice-flow rate was metered by a bank of four calibrated rotameters in parallel, each followed by a control valve. Each rotameter was calibrated over the entire range of flow by using a weigh tank. Calibrations were checked periodically. The metering range available was 200 to 40,000 lb per hr. From the control valve, the oil passed through the header and the orifice and returned to the storage tank. Auxiliary heating and cooling of the oil was done in an independent parallel system connected to the storage tank. This system consisted of a centrifugal pump and a heat exchanger with steam and cold-water connections.

The water-pumping system was constructed of $1^1/_2$-in. brass pipe with bronze fittings. The water flowed from an overhead surge tank, which served both to maintain a constant head and to remove entrained air, to a centrifugal pump. After the pump, the stream passed through a calibrated 10 to 100-gpm guided-float rotameter and control valve and into the header and orifice. From the orifice, the flow passed through a double-pipe heat exchanger and into the surge tank.

The pressure difference across the orifice was measured on differential manometers using either test fluid (oil or water) over mercury, air over test fluid in a vertically mounted manometer, or air over test fluid in a slanted manometer with a slope of 0.275. The connecting tubing was kept horizontal in regions where there might be fluid-density differences in the line.

Point temperatures were measured by single copper-constantan thermocouples inserted into the stream. The temperatures were indicated on a self-balancing electronic potentiometer. The measured temperatures are believed to be within ± 0.3 deg F of the true temperatures.

The test fluids were water, Gulf 896 oil (viscosity 1.86 cp at 150 F), and modified Gulf Crown C oil (viscosity 63.0 cp at 150 F).

Experimental Procedure. For the oil runs, the oil was brought up to the desired operating temperature by circulation through the auxiliary system. The main oil pump was then started with the by-pass full open and the system vented under pressure to remove trapped air. The valves were adjusted to give an orifice-flow rate near the desired maximum and the manometers drained until there was no further evidence of air in the lines. The drain lines were closed, the flow rate adjusted to the final value, and the system allowed to run to equilibrium. For each run, the data taken were: Rotameter number and reading, tank temperature, inlet fluid temperature, and manometer type and reading. Manometer readings were taken at intervals of 3 to 5 min, and three essentially identical readings were considered necessary to indicate attainment of equilibrium. When the run was completed, the flow rate was adjusted for the next run and the discharge valve adjusted to maintain a small positive pressure on the shell.

The minimum acceptable reading on any manometer was a 1-in. differential; for smaller pressure differentials, a manometer of greater sensitivity was used. At the completion of a series of runs, "no-flow readings" of the manometers were taken. If the manometer levels were within ± 0.025 in. of exact balance, the deviation from the null point was applied to the indicated differential as a correction for minute quantities of air in the lines or slight misalignment of the equipment. "No-flow readings" greater than the 0.05 maximum differential mentioned were considered to indicate appreciable air in the manometer lines. The series of runs was rejected and the condition corrected.

For low flow rates of Gulf Crown C, the rotameters were very sensitive to small changes in temperature, and it was necessary to weigh the flow for each run.

For water runs, the procedure was similar to that for oil runs except that cooling was accomplished by direct introduction of cold water from the mains and by-passing was not required since a centrifugal pump was used.

Calculations and Presentation of Experimental Results. An overall orifice coefficient was calculated for each experimental run, using Equation [1]. An orifice Reynolds number also was calculated for each run, using Equation [2]. For each orifice, the results were plotted as C versus Re. The results for orifices 5.02-S, $(D - d) = 0.2290$ in., $L = 0.0135$ in.; and 5.20-1, $(D - d) = 0.0537$ in., $L = 0.8934$ in., are shown in Figs. 5 and 6.

The orifice length-to-width ratio Z defined by Equation [3] was found to be important. For all of the orifices the experimental results are summarized in Figs. 7 and 8, where the coefficient is plotted as a function of Z with the orifice Reynolds number as a parameter. If more detailed information is desired, the tables of experimental data and plots of C versus Re for twenty-one orifices may be obtained from the American Documentation Institute.[4]

Experimental Errors. The calculated orifice coefficient may be in error as a result of errors in the measured value of the diameters, the flow rate, and the manometer reading. The Reynolds number may be in error as a result of errors in the flow rate, the fluid temperature, and the measurement of the fluid viscosity. The maximum possible error in the coefficient was calculated to be in the range from 1.8 to 5.2 per cent, and that for the Reynolds number from 0.9 to 2.1 per cent. It is believed that these maximum errors were never obtained and that the orifice coefficients are within ± 2 per cent of the true value and the Reynolds numbers are within ± 1 per cent.

For concentric orifices, there is also a possible error due to imperfect centering of the disk in the shell. This error, which can be quite significant in the viscous-flow regime, was discussed earlier in the comparison of theoretical and experimental results.

[4] The tables and plots have been deposited as Document No. 4993 with the ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress, Washington 25, D. C. A copy may be secured by citing the Document number and by remitting $6.25 for photoprints, or $2.50 for 35 mm microfilm. Advance payment is required. Make checks or money orders payable to: Chief, Photoduplication Service, Library of Congress.

How RF Concerns the Wood Industry

By J. W. MANN,[1] TACOMA, WASH.

High-frequency alternating currents to the order of several millions of cycles per second applied to dielectric materials are referred to as "RF heating," and are used widely in wood lamination as a means of setting synthetic adhesives in such members. Aspects of federal rules and regulations and the needs of industry are set forth. Means of accomplishing industrial applications; methods of estimating time cycles; the unique properties of the high-frequency field of force such as selectivity to conductive paths and nonuniformity of the heat placement in dielectrics are covered in some detail with specific examples.

Introduction

THE wood industries are constantly encountering a problem of procuring clear lumber for manufacturing into end products. Decreasingly available supplies of old-growth timber have forced users to employ timber of poorer grades and better, and more fully, to utilize those high grades which are still available to them.

This is illustrated well by the search of the plywood industry for better and more economical methods of utilizing lower-grade peeler logs—to edge-glue veneer from which defects have been clipped, and to patch finished panels showing defects. As logs become scarcer, this search will be intensified. The growth of dielectric high-frequency heating in the wood industries is a direct result of this search for better methods of conservation and utilization of the raw materials currently available. This intensification has been substantial since the end of World War II, and continues apace today without prospect of let up. It is upon the economic premise that the equipment produced by electronics manufacturers substantially assists industry in its search for better utilization of its available raw materials and assists industry in employing materials not otherwise usable, that the dielectric high-frequency heating industry has grown and prospered.

Among wood end products produced with dielectric high-frequency heating equipment are the following: Door stiles and rails; piano sounding boards; spruce piano components; chair legs, seats and arms; drawer runners, faces and sides; table and desk tops, skirts and siding; stadium seats, barn and church rafters; structural beams, stringers and boat keels; oak mine-sweeper parts, planking, and hulls; boat parts; finger-joined end-glued lumber; shingle panels; siding; pipe and tank staves; tent poles; radio masts; hardwood-veneer sheets; Douglas fir, endless, edge-glued veneer; TV and radio cabinets; shelving and millwork parts; store display cabinets; edge-glued boards and panels; furniture core stock and numerous others.

Types of equipment produced for separate uses are so numerous that unnecessary length would be added to this brief if all such different applications were described in detail. Prior to 1950 batch-production electrodes and clamping devices dominated the means of applying the high-frequency field of force to industrial uses, while the more recent trend in large equipment is to continuous-process handling in which especially striking progress has been made.

The abbreviation RF concerns the whole subject of heating dielectrics by means of high-frequency alternating electromagnetic fields of force, mostly in the range of frequencies above 3 megacycles (mc).

Many dielectrics or semidielectrics, besides wood and certain types of glues, are being heat-treated, dehydrated, preheated internally, and fabricated by the application of RF.

In the present state of the art, high-frequency alternating current, to the order of megacycles per second is obtained most conveniently from conversion of conventional 60-cycle-per-sec (cps) power sources through the medium of three-element vacuum-tube oscillators. Incidentally, the capital investment per kilowatt of RF output power for high-frequency generators, 5 kw and above, ranges around $600 per kw.

The electron swing of an alternating current is quite analogous to the motion of a swinging pendulum. Frequency is a measure of the number of complete cycles occurring per second.

Theoretically speaking, high-frequency alternating currents are only different from slow-frequency alternating currents with respect to the degree to which any one phenomenon manifests itself. Sixty-cycle alternating electromagnetic fields show negligible heating effects in dielectrics as compared to those in the millions of cycles-per-second range. In many cases, other factors remaining unchanged, the heating effect is proportional to the frequency. Therefore high frequencies seem desirable for dielectric heating and processing purposes.

Frequency; Wave Length; Resonance

Alternating electromagnetic fields of force are propagated through space at the speed of light; thus one complete cycle may be expressed in wave length, the wave length being inversely proportional to the frequency.

At a frequency of 60 cps the wave length is to the order of 3100 miles. Experience indicates frequencies 3 mc per sec and above to be most adaptable for dielectric heating. At 3 mc the wave length in space is approximately equal to 100 m or 327 ft.

Wave length involves load dimensions which in turn involves frequency response which again requires that "line length" be defined. The communications radio engineer is primarily concerned with radiating a maximum of electromagnetic energy into space by means of an antenna or line at the expense of the source, whereas the engineer concerned with RF industrial heating obviously bends every effort to minimize radiation into space and, instead, to convert a maximum of electrical energy into heat energy within the bounds of the dielectric being processed, the latter included as a part of a line length of the applicator. The communications engineer deals with loads in the form of antennas which are stable and simple with respect to geometrical shapes as compared to the problems met with in RF dielectric heating. In the latter, wood and plastics are especially variable, changing electrical constants with every second of time during the heating cycle.

The geometrical shapes of clamping means in three dimensions can be infinite in variety and the dielectric may be held stationary in the field or continuously moved through the system as may be desired.

---

[1] Vice-President, Mann-Russell Electronics.

Contributed by the Wood Industries Division and presented at the Spring Meeting, Portland, Ore., March 18–21, 1956, of The American Society of Mechanical Engineers.

Note: Statements and opinions advanced in papers are to be understood as individual expressions of their authors and not those of the Society. Manuscript received at ASME Headquarters, January 19, 1956. Paper No. 56—S-17.

The communications engineer normally cuts his radiating antennas to correspond closely to quarter wave lengths or multiples of quarter wave lengths corresponding to the frequency employed. Such a resonant length has been designated as the "electrical length" of the system regardless of geometrical shape, meaning that its particular fundamental frequency response is identical with the corresponding calculated space wave length. Dielectric loads are also tailored to conform to some desired electrical length.

It is significant to note that high-frequency electromagnetic energy is generally transmitted or received by exciting the source and the sink into a state of oscillation involving reflection of energy which results in so-called "standing waves." This condition in turn requires that the line lengths correspond to the calculated space wave lengths. Thus there is close correlation in RF heating between the geometry of a dielectric load and its inherent electrical properties, such as dielectric constant and self-inductance thereof, together with the inductance and capacity of the associated connections, electrode applicators, supporting framework, shielding, and so on.

Just as in communications, the source of the electromagnetic energy may be transmitted to the sink through the medium of a nonresonant, nonradiating, traveling-wave transmission line. However, owing to the changing characteristics of a dielectric load during the heating cycle, such a linkage is highly unsatisfactory in many types of application to dielectrics.

A superior means of furnishing energy from an electronic generator to an electrode configuration which applies the field of force to the dielectric constitutes, in effect, making the load configuration part of the oscillator itself and eliminating any link between the two, such as is described as a nonresonant traveling-wave transmission line.

Generators of the type which encompass the work load as a part of the oscillatory circuit rather than those connected to the load by output coupling means are usually self-excited generators. They have no frequency-control circuit such as a crystal oscillator to maintain fixed frequency. Self-excited generators have been found adaptable to work with wood whose dielectric swing takes place on heating of the load. During heating, therefore, the *LC* ratio of the oscillatory circuitry shifts, causing a shift in oscillating frequency; at the 13.56-mc frequency as much as plus and minus 250 kilocycles (kc) from center frequency at the 27.12-mc frequency up to 500 kc, and at the 40.68-mc frequency up to 700 kc but many variations are less, depending on electrode configuration, load, and so on.

Configurations and types of dielectric loads which have relatively slow natural resonant periods are best suited to being excited by means of correspondingly long wave length (slow frequency) sources of RF energy.

Following are some practical examples of load dimensions to frequency. In loads of large dimension (i.e., 10-ft-long veneer for edge-gluing), a frequency of 13.56 mc is most readily employed; in loads of medium size (i.e., 4-ft-long glued shingle panels), a frequency of 27.12 mc is a natural; while on smaller loads (i.e., plastic sealing and butt finger joints), a 40.68-mc frequency makes the best adaptation in circuitry which permits the load to function as an integral part of the generator oscillatory circuit.

Radiation and Regulation

It is difficult to eliminate radiation of RF energy into space from any high-frequency source such as an RF industrial installation and hence the Federal Communications Commission has become interested in interference to communications which might result from improper installations.

In 1945 rules were promulgated (Rule 18) and space was allotted in the spectrum for industrial applications of RF and, fortunately, whether by design or by expediency, frequencies were allotted which later proved to be of wave lengths quite suitable to matching the dimension of many industrial dielectric applications; three specific examples of this were just pointed out.

Little was understood at the time concerning RF applications to industry and, as a result, only these few and meager band widths were allotted in an overcrowded spectrum already practically monopolized by communications.

These three frequencies have tolerance plus and minus which are inadequate if industrial, scientific, and medical equipment is to confine its operation to "on-band" use. Restriction for "off-band" use are so tight that many applications which might be used cannot currently be considered. To be practical and of use to industry the present bands must be widened for on-band operation with harmonic and near field protection provided for the frequencies employed for dielectric high-frequency heating applications.

It will be noted that no frequency is provided by Rule 18 for on-band operation in the range approximating one half of the 13.56-mc band. As pointed out the load size is one factor which determines proper operating frequency, and time cycle required is a second and of course the most governing economic factor. It would appear impractical to make a 16-ft lumber edge gluer operating on 13.56 mc and to develop at such frequency the required power. If a frequency were available it undoubtedly would be employed. The swing of a self-excited oscillator operating on a large wood load at such frequency might be confined to less than 100 kc each side of center frequency.

Use of radio frequency by industry is rapidly becoming indispensable because of two unique properties of RF covered in further detail in the following: (1) Quality of selectivity, and (2) internal heat placement of the field. Most, if not all dielectrics, not only class as electrical insulating media but also as very poor conductors of heat. Consequently, conventional methods of imparting heat to the interior portions of dielectrics, such as wood or plastics especially, result in time-consuming and inefficient procedures. A block of wood may be charred on the surface by conduction heat applied externally without appreciably affecting the temperature farthest from the surfaces but the same block of wood exposed to an RF field may be charred internally before the surface indicates what is taking place within. By applying exterior heat and humidity control simultaneously with RF internal heating, the time required for kiln-drying lumber may be cut substantially.

Quality of Selectivity

RF exhibits a unique selective effect which, in the majority of applications, is basic to many processes. For instance, a wet glue line if positioned in the electric field, substantially parallel to the electric lines of force, will absorb energy at a rate equal to about ten times the rate at which the adjacent wood absorbs energy. The electric flux fringes into the glue line in preference to the body of the wood. This is a patented process known as "parallel bonding," basic especially to edge-bonded panels, lumber, veneers, and the like.

In the drying of veneers, high moisture-content pockets may persist. If such veneers are subjected to the selective effect of RF the mc of the panels may be equalized completely. Incidentally, this selective effect has been found to be of potential value in the pasteurization of wine and the killing of infestation in cereals, and so on.

It is now conceded that RF is indispensable where heat must be generated deep in a dielectric or where certain portions or par-

ticles of a nonhomogeneous substance must be differentiated between by means of the selective effect.

It may be stated here that certain glues, such as animal glue, although selective due to mc, are not suitable where RF is concerned. Certain synthetic glues are highly selective and polymerize rapidly with rise in temperature, resulting in a bond which is waterproof or highly so, with properties comparable to the strength of the natural wood itself.

Nonuniformity of the Field

Theoretical consideration has established the premise that a dielectric placed between the plates of a condenser and subjected to RF power will be heated throughout absolutely uniformly. This can only be true provided the electromagnetic waves sweep one after the other, through the dielectric in the form of a traveling wave. Thus the effect would be average and result in uniformity.

Barring the losses of heat resulting from surface radiation, convection, and conduction, experience strongly indicates a nonuniform placement of energy concentrated on either side of the center plane located half way between the plates of the condenser encompassing the dielectric load. This has been conveniently designated as the "two-spot" effect and the regions of intensified heat energy as "energy concentrations." It is significant to observe that in a "single-ender" or quarter-wave system where one condenser plate opposes a grounded plate, only one highly heated concentration will be observed somewhere between ground and the live element. On the other hand, in a "double-ender" system, two concentrations will be observed as just described.

This contradiction of theory requires some deep study. As a matter of fact, RF functions best where standing waves of energy are propagated across the dielectric. Each half cycle of alternating current, where current and voltage are substantially 90 deg out of phase, contains positive and negative loops of energy, the peaks of which are located at or near one-eighth wave intervals. Since the wave "stands," as it were, with respect to the dielectric, so also must the peak-of-energy-conversion regions stand in the dielectric. As a result the center portion of each quarter wave will be subjected repeatedly to a pulse of heat energy and result in the concentration-of-heating effect.

Of course the contour of electrodes also may influence the fringing of electric flux into the dielectric and result in hot spots not related.

Harmonic content of the fundamental field also may be a contributor to the nonuniform placement of heat energy in a dielectric.

Naturally, nonuniform heating may or may not be desirable in certain processes. However, knowing the causes, considerable correction may be brought about and the effect employed to advantage in dielectric high-frequency heating.

Because of the peculiar phenomena of energy placement just described, the RF engineer is faced with the problem not only of staying closely within some assigned frequency and related wave length, but also he must establish a standing-wave pattern or node which will locate the energy concentration in the body of the dielectric so that certain maximum desired effects may ensue in the right places. In order to function, the standing-wave pattern of nodes and antinodes must bridge all discontinuities between generator and load, such as variable condensers and inductances used for tuning each quarter wave length of the system.

Assume wood in the form of an edge-glued package and clamped edgewise and confined to a flat plane between plates which also function as the opposing plates of a condenser.

The frequency response of such a package is affected and controlled by the following factors:

1 Capacity of such condenser varies directly as the area of the dielectric covering the condenser plates.

2 Varies inversely as the thickness of the dielectric.

3 Varies directly with the several composite dielectric constants of substances composing the dielectric package; i.e., wood, glue, moisture, etc.

Estimate of Time Cycles

By experience, the Btu output required to do a certain job may be closely estimated, based upon the observed mass rise in temperature of types of woods and glues which produces consistent bonds. In "parallel bonding" it may be assumed safely that nearly all the mc of the glue will be evaporated. Hence, by adding the Btu content of the wood to the Btu stored in the mc of the wood plus the heat of vaporization required to evaporate the mc in the glue, one may calculate from these data the approximate time cycle to be expected from any given-size generator.

Matching Load and Generator

Preferably, the more nearly the load becomes a part of the generator standing-wave pattern, rising and falling in unison and in the same phase, the more stable the whole system becomes, as previously described.

Some systems use what is known as loose coupling, which method pertains almost exclusively in communications. For dielectric loads, if an attempt is made to couple too closely for the purpose of using the power source to capacity, the load may suddenly rob the oscillator of its stored energy, oscillations cease, and the result is a heavy short circuit and power outage. The standing-wave pattern of the load is 180 deg out of phase with that of the generator. Such a system is quite unstable when loaded with widely changing loads as are wood and glue.

Single-ender systems may be described as those in which the load spans a quarter-wave line length of electromagnetic energy usually recognizable by the load being interposed between a "live" plate and the grounded framework of the press.

In a double-ended system the load is placed between plates of opposite instantaneous charge, the press frame being symmetrical to, and neutral to, both plates. This system is to be preferred in most cases because of more nearly uniform distribution of heat and balanced charges.

The shape and size of electrode configurations must be designed to fit the work in hand. They may be shaped to spot set glues, to form Z-patterns or a partial set pattern produced by stray field and split-pole methods. By these means, with a given power source, production rates per kilowatthour may be raised, the unset glue being later polymerized, more or less, by the residual heat stored in the wood. In most of these cases parallel bonding is employed. In the case of scarfed joints, perpendicular bonding has been found a preferred method at a sacrifice of time cycle, thus producing near full strength immediately and solid-setting feather edges of the scarf.

Wherever possible, continuous-process bonding devices should be used because of the higher production rates possible per kilowatt as compared to the stop-and-go system.

Other Factors

Since the press frame work, auxiliary parts together with the electrode applicator, form a part of the whole radio circuit, mechanical design of the press is governed by the physical properties of electromagnetic waves, electromagnetic fields and electric charges, resonant line lengths, and so on.

The size, strength, and design factors of clamping devices are governed by the character of the load and the practicability of suiting certain wave lengths and wave patterns to the load.

Hence the RF engineer and the mechanical engineer must come to mutual understanding in the complex design of RF heating equipment.

Comment should be made with regard to preparation of stock and the control of moisture contents because RF involves precise factors which in turn are best served by precise woodworking methods, good workmanship, and control.

Glues which behave best in the RF field are synthetic and are of two types: highly water-resistant and fully waterproof. The ureas are highly water-resistant and polymerize at relatively low temperatures while the melamines, the resorcins, and the phenols are waterproof but require higher temperatures to set and result in comparatively longer cure-time cycles, other factors being equal.

The amount of pressure required for most fabricated wood parts ranges between 50 and 150 psi and may be supplied through the medium of air or by hydraulic means. Air pressure is preferred because of superior speed and cleanliness as compared to oil-pressure systems.

Other factors remaining the same, dense woods normally absorb more energy in a given time than light porous woods. Edge-gluing results vary between wood of differing types of grain structure, such as vertical or slash, summer or winter, and so on. These factors sometimes require adjustments in intensity and disposition of the RF field to avoid pitfalls of too great capacity variation and gluing results.

As time elapses, more and more use is being made of RF in the wood industry as the need for conservation develops ever-increasing economic pressure.

These needs are being met by the manufacture of RF equipment and even anticipated with new equipment, one example of which is a new configuration for setting end-glued finger joints continuously at high rates of lineal speed.

Economics

A few of the technical aspects of dielectric high-frequency heating have been covered, but each has its specific relation to the economics of production in the wood industries. Were the savings obtained, or the betterments in production processes in doubt, industry would not turn as it has recently to the use of this new tool of production. To be specific, a door factory had a problem of waste; so great was the trim from ends of dried and otherwise good wood, that raw-material costs exceeded that which could be borne in competition with lower-cost producers. Installation of an electronic edge-gluing machine made possible utilization of the waste product in laminated built-up stiles; short pieces between outer bands of solid wood with veneered faces. Such were the savings that more doors were produced with the same mill input of wood, utilization increased, and annual savings in wood cost exceeded the value of the electronic unit by over ten times.

Not only do the lower costs of production work out to the financial betterment of the user but another factor deserves consideration. This factor is the ability of a wood-using industry to make products not otherwise possible of manufacture. This is illustrated by a producer of vertical-grain, diagonally edge-glued piano sounding boards the production of which was impractical before application of RF to the problem. Also built around electronic edge-gluing is an entire department of a wood-pipe and tank manufacturer who found it possible to utilize very low-cost wood scraps and end cuttings as raw materials and so reduced cost of production that wood pipe became competitive with stainless steel and iron pipe for certain paper, chemical, and corrosive-materials use.

Simple to illustrate is the competitive situation in edge-glued lumber panels for producers of TV cabinets, sideboards, kitchen cabinets, and myriads of like products; hand-glued batch production was so high in cost that substitutes in the form of low-grade plywood were coming to the market in increasing quantities. Installation of continuous-process lumber edge gluers with automatic cutoff saws for the production of these panels has shown ability to produce over 25,000 fbm in an 8-hr shift—placing the producer right back in the thick of the competitive scrap with lower-cost substitute materials, thus not only retaining but expanding his market at a rate of return more favorable than previously he could expect to receive.

Installation of electronic, continuous process, veneer edge-gluing equipment in plywood plants has expanded greatly in recent years showing greater utilization of random-width veneers, less rejects to the burners, and increased veneer output per log of raw material peeled. Elimination of tape in splicing, sanding of panels, and ability to patch exterior-grade plywood electronically has added to the production and lowered the cost of panels for most of the progressive plywood mills in the California, Oregon, Washington, and British Columbia region of plywood production. Low operating and maintenance cost is a factor favoring the economics of production through use of dielectric high-frequency heating.

Effect of Ambient and Fuel Pressure on Nozzle Spray Angle

By S. M. DE CORSO[1] and G. A. KEMENY,[1] E. PITTSBURGH, PA.

Diametral samples across the fuel spray at a distance of $4^1/_2$ in. from the nozzle tip were obtained for ten centrifugal-type nozzles of 9 to 100-gph capacity, having nominal spray angles of 45 and 80 deg. The data were taken over a fuel-pressure range of 25 to 400 psi and for ambient pressures from 0.1 to 8 atm. These diametral spray distributions were reduced to equivalent spray-angle values which when plotted against ambient and fuel pressure provided a summary of the pressure effects on the spray angle. It was found that the spray angle decreased markedly with increasing fuel and ambient pressure. An explanation of the phenomenon is given. The equivalent spray angle was found to be a function of the product of fuel pressure drop and ambient gas density to the 1.6 power, i.e., $P\gamma^{1.6}$.

Introduction

IT HAS been found that the fuel-nozzle spray angle is an important factor in the burning of liquid fuels in gas-turbine combustors. Combustion efficiency, flame length, and combustor-wall temperatures are all affected by the nozzle-spray distribution.

This paper deals with the effect of ambient and fuel pressure on the spray distribution of centrifugal nozzles. The centrifugal fuel nozzle is an important type, widely used in industrial gas turbines and in turbojet engines. This type of nozzle is designed to give, at atmospheric pressure, a spray of approximately "hollow-cone" shape which gives optimum atomization for gas-turbine applications.

A typical design is shown in Fig. 1, together with a spray shape obtained at atmospheric pressure. The essential elements of the nozzle are the swirl slots, swirl chamber, and outlet orifice. The fuel, under pressure, passes through the swirl slots into the swirl chamber where a vortex is created. The fuel then issues from the outer orifice with a velocity whose axial and tangential components determine the spray-cone angle. The spray-cone angle is defined as the angle between the tangents to the spray envelope at the nozzle. Hereafter in this paper the term spray-cone angle will refer to the latter angle. Actually, the surface of the spray near the nozzle tip forms a hyperboloid, but for all practical purposes, this may be considered a cone at low ambient gas pressure. Detailed discussions of the fuel flow in nozzles may be found in references (1) and (2).[2]

A particular nozzle is identified by giving the flow rate at a certain pressure drop in terms of some fuel type and by giving the spray-cone angle. For example, a 60-gph 80-deg nozzle is one which will discharge 60 gph of diesel fuel at 100-psi nozzle pressure drop and has a spray-cone angle of 80 deg at atmospheric pressure.

[1] Westinghouse Electric Corporation.

[2] Numbers in parentheses refer to the Bibliography at the end of the paper.

Presented at the Gas Turbine Power Division Conference, Washington, D. C., April 16–18, 1956, of The American Society of Mechanical Engineers.

Note: Statements and opinions advanced in papers are to be understood as individual expressions of their authors and not those of the Society. Manuscript received at ASME Headquarters, January 16, 1956. Paper No. 56—GTP-3.

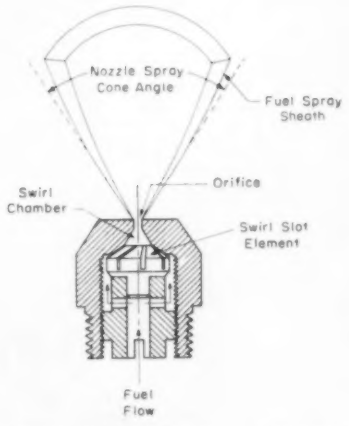


Fig. 1 Typical Centrifugal Nozzle

Joachim and Beardsley in 1927 (3, 4), and Joachim in 1927 (5), have investigated centrifugal sprays in connection with diesel-engine injectors. They worked at nozzle pressure drops of 6000 to 10,000 psi and ambient gas pressures of 200 to 600 psig. These pressure levels are of course much higher than those generally in use in gas turbines today. From photographs of the sprays, they show a decrease in spray angle with increasing ambient gas density. D. W. Lee (6) in 1935, in similar experiments on diesel injectors, also found a decrease in spray angle with increasing gas density. In their diesel-injector experiments, the spray outline was photographed during the time interval from spray initiation to 0.006 sec later. Watson and Clarke (7) in a gas-turbine combustion study have noted the decrease in spray angle with increasing gas density, which they illustrated by a photograph. Others (1, 8) have noted a slight decrease in spray angle with increasing nozzle pressure drop at atmospheric pressure. A decrease in angle with increasing air pressure drop was described by one of the authors (9) for a swirling annular air jet such as is used in connection with an air-assisted fuel nozzle.

The data presented here cover nozzle pressure drops of 25 to 400 psi and ambient gas pressures of 1.5 to 114 psia. The nozzles were sprayed into quiescent gas at room temperature. The data were obtained principally from tests on ten nozzles of commercial type which were obtained from a nozzle manufacturer. The nozzle capacities (in gallons per hr of diesel fuel at 100-psi pressure drop) were 9, 20, 45, 60, and 100 gph with spray angles of 45 and 80 deg in each capacity, i.e., ten nozzles in all.

The fuel used in all the spray tests was diesel oil with the following properties at 77 F: Specific gravity, 0.84; surface tension, 30.7 dynes/cm; kinematic viscosity, 2.43 centistokes.

Description of Apparatus and Procedure

The test facilities include the high-pressure tank, instrument panel, and associated fuel pumps, and controls. Fig. 2 shows the 18-in-diam high-pressure tank; Fig. 3, the control panel; and Fig. 4 is a flow diagram showing the major components of the

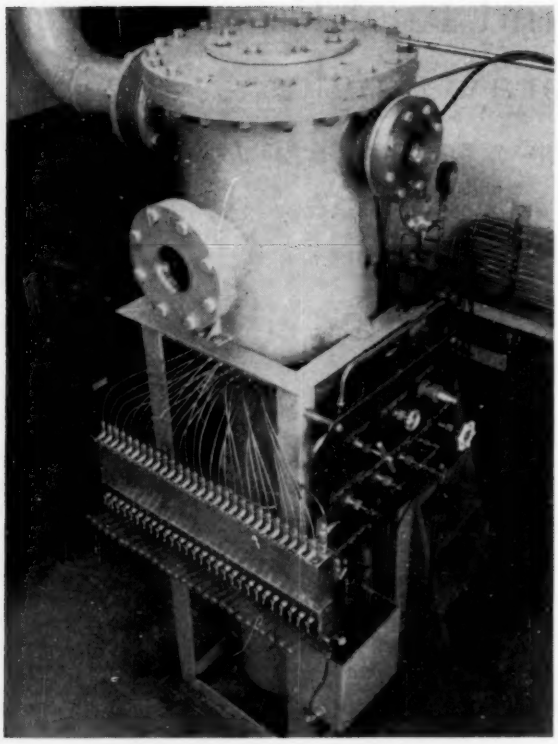


FIG. 2 HIGH-PRESSURE SPRAY TANK

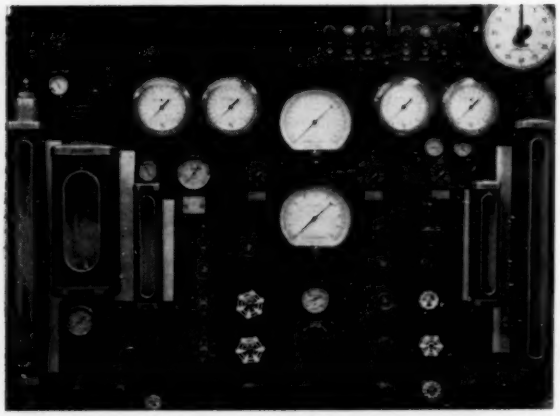


FIG. 3 CONTROL PANEL

nozzle-testing installation. As safety precautions, the following measures were taken: (*a*) Nitrogen was used as ambient gas for the fuel spray. (*b*) The tank with observation windows and sampling equipment in place was tested hydrostatically at 200 psig. (The maximum operating pressure anticipated was 100 psig.) (*c*) A pressure-relief valve set at 105 psig was provided and the tank also was equipped with a 6-in. 150-psi blowout diaphragm which is shown in the top left corner of Fig. 2. If an increase in tank pressure above 150 psig should occur, the diaphragm would rupture, and the contents of the tank would discharge through a 6-in. pipe to the outside of the building.

Fig. 3 shows the control panel which includes the gages and valves required for operating both the new pressure tank and the atmospheric-pressure spray system. To avoid confusion during tests, the valves were color coded as follows: Red for those used only for the high-pressure tank; white for those used only for the atmospheric tank; and black for valves which are common to both systems. In order to reduce the number of pressure gages required, lines were run from pressure-measuring stations to a gage manifold. Thus any available gage can be connected to any pressure station as required. A few of the gages, however, are permanently connected to pressure-measuring stations.

Fig. 4 shows the controls which are required for operating the two nozzle-testing tanks from a single fuel-pumping system. The two tanks each have their own fuel reservoir, and the low-pressure pump by pass fuel is always returned to the tank from which fuel is being pumped; the fuel flowrators meter at the pressure level existing between the low-pressure and high-pressure pumps. A pressure-relief valve limits the maximum pressure here to 200 psig to prevent damage to the flowrator tubes. Because the flowrators meter fuel at a point upstream of the high-pressure pump, a water-cooled by pass line from this pump discharges back into the high-pressure pump inlet. The fuel system is adequate for testing nozzles up to 100-gph capacity at pressures up to 400 psi.

The air system is used for supplying compressed air from a house compressor for air-atomizing nozzles sprayed into the atmospheric tank. Nitrogen is used for pressurizing the high-pressure tank as well as for the atomizing gas when air-atomizing nozzles are to be tested at elevated ambient pressures. The nitrogen exhausted from the tank passes through a mist separator which can be bled continuously to a fuel container. Flow through the exhaust line can be throttled by a diaphragm valve which is operated by a manual control located at the control panel.

The spray-sampling system has 28 radial sampling tubes spaced 5 deg apart on the arc of a $4^1/_2$-in-radius circle with the nozzle tip located at the center of curvature of the arc. The $^3/_{16}$-in-diam tubes are drilled out to 0.168 in. diam at the sampling tip. Each tube, as seen in Fig. 2, passes through a packing gland to the outside of the tank. The spray-sampling procedure can best be explained by looking at a cross section of the brass bar containing the reservoirs, as shown in Fig. 4. While spray is being collected, the top 28 toggle valves are open, the bottom 28 toggle valves are closed, valve 18 H is closed, and the manifold return valve 17 H is open. This allows fuel to collect in the 28 separate fuel-collecting reservoirs, each reservoir being a hole drilled in the block with a toggle valve at its inlet and outlet. Each reservoir is also connected by a slanting hole to a manifold which, in turn, is connected to the tank through valve 17 H. This allows fuel to flow freely through each sampling tube to its reservoir. After spray-sampling has been stopped, the top toggle valves and 17 H are closed. All reservoirs are then reduced to atmospheric pressure by opening 18 H. At this point, the bottom toggle valves are opened and fuel-spray distribution can be read directly from the graduated test tubes.

If any reservoir overflows during the sampling, the corresponding test tube will be filled above scale and a test of shorter duration must then be made. The overflow fuel collects in the bottom of the manifold and can be bled out into a small auxiliary fuel container by opening valve 18 H. After readings are taken, all the tubes can be drained into the auxiliary fuel container as is being done in Fig. 2. Fuel from the auxiliary tank can be pumped back into the high-pressure tank using the regular pumping system.

Sampling is started and stopped by movement of a shutter which covers the sampling tubes. The shutter mechanism is operated by nitrogen pressure. The timer switch shown in the upper right-hand corner of the control panel automatically opens the shutter for the time interval desired.

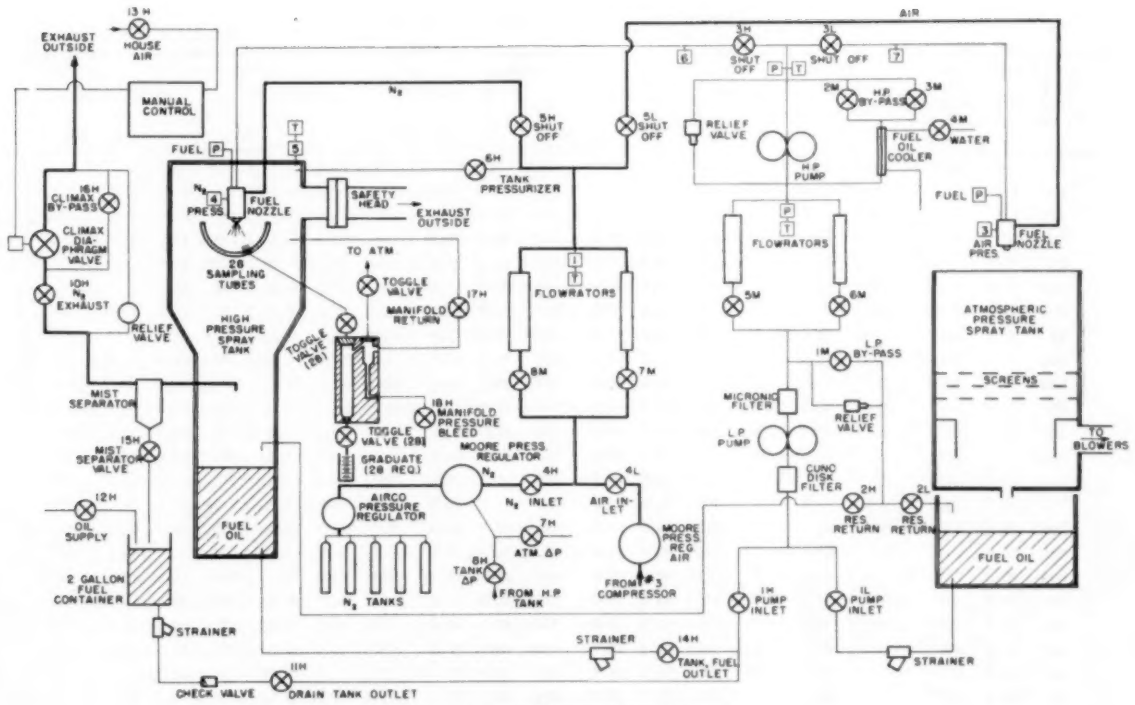


FIG. 4 NOZZLE-TESTING FLOW DIAGRAM

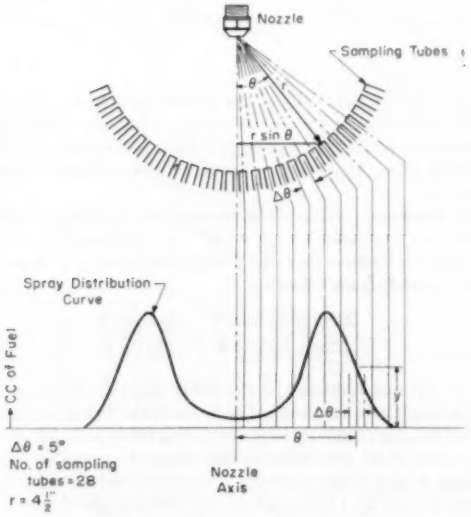


FIG. 5 SPRAY-SAMPLING TUBE ARRANGEMENT

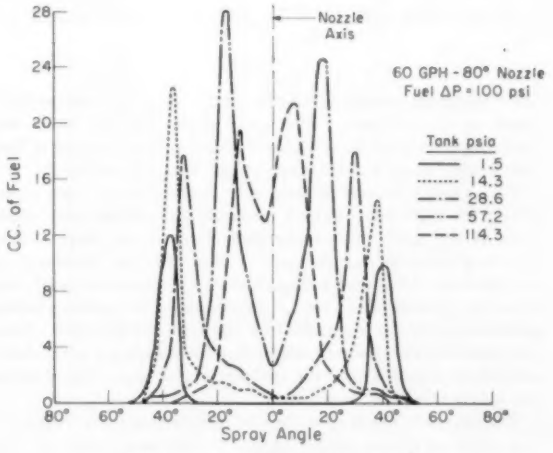


FIG. 6 FUEL-DISTRIBUTION CURVES

It was found that spray-distribution data could be obtained very rapidly with this system. For example, a complete set of data on one nozzle (20 spray distributions) can be recorded in about 3 hr.

Table 1 gives the orifice diameters of all the nozzles.

TABLE 1 NOZZLE-ORIFICE DIAMETERS

Capacity, gph	9	20	45	60	100
45-deg angle, in.	0.0332	0.0585	0.0757	0.0830	0.1046
80-deg angle, in.	0.0315	0.0554	0.0779	0.0888	0.1162

DISCUSSION OF TEST RESULTS

The spray-cone angle of a fuel nozzle does not give an accurate description of the spray distribution unless the spray sheath is very thin, which is generally not the case. Spray distributions were obtained in these tests by taking fuel samples at points along an arc as shown in Fig. 5, the samples being proportional to the fuel flow at the respective sampling points. This produces spray-distribution curves such as those shown in Fig. 6, where the ordinates of the curves represent the amount of fuel collected at

TABLE 2 SPRAY-DISTRIBUTION DATA

TEST CONDITIONS				NOZZLE																	
Nozzle Press. Drop P, psi	Tank psia	Gas Density γ lb/ft.^{3}	$P\gamma^{1.6}$	100 GPH - 80°			60 GPH - 80°			45 GPH - 80°			20 GPH 80°			9 GPH 80°		
				ψ	φ	φ/φ_o	ψ	φ	φ/φ_o	ψ	φ	φ/φ_o	ψ	φ	φ/φ_o	ψ	φ	φ/φ_o
25	1.5	.00732	.00963	83.5	82.5	1.140	81.0	82.0	1.163	80.5	81.7	1.153*	76.0	77.0	1.100*	68.5	71.7	1.185*
100	1.5	.00732	.0385	80.0	77.0	1.061	79.5	79.5	1.129	78.5	79.2	1.120	82.0	79.7	1.138	69.1	71 6	1.185
225	1.5	.00732	.0866	----	----	-----	79.2	78.2	1.110	77.5	82.1	1.160	82.0	81.7	1.167	61.5	65.2	1.078
400	1.5	.00732	.1540	----	----	-----	79.5	80.3	1.140	76.5	76.7	1.083	82.0	82.3	1.175	53.9	60.9	1.005
25	3.58	.01748	.0375	82.7	83.0	1.146	78.5	80.2	1.138	77.3	79.2	1.120*	73.6	74.8	1.069*	66.3	69.2	1.142*
100	3.58	.01748	.150	75.5	76.2	1.051	77.5	77.7	1.102	77.3	77.8	1.100	79.0	78.2	1.117	65.8	68.8	1.137
400	3.58	.01748	.600	----	----	-----	68.5	72.0	1.022	69.2	71.7	1.013	74.2	78.2	1.117	51.0	57.2	0.945
25	7.15	.0349	.1176	82.4	83.0	1.146	75.2	78.2	1.110	73.3	75.8	1.070*	70.0	72.0	1.028*	63.2	67.1	1.110*
100	7.15	.0349	.470	75.1	75.0	1.034	73.7	75.5	1.070	74.2	75.5	1.067	73.0	76.0	1.086	61.2	65.2	1.079
400	7.15	.0349	1.880	----	----	-----	56.2	62.5	.887	58.6	64.0	.904	62.7	69.8	.997	41.8	51.3	.848
25	14.3	.0698	.350	81.0	81.5	1.125	71.5	76.0	1.080	67.5	71.8	1.015*	64.0	68.5	.978*	58.8	64.9	1.071*
100	14.3	.0698	1.40	73.5	72.5	1.000	66.0	70.5	1.000	67.5	70.8	1.000	63.5	70.0	1.000	53.8	60.5	1.000
225	14.3	.0698	3.15	----	----	-----	57.0	61.0	.865	61.0	62.7	.885	55.0	63.2	.903	40.5	49.0	810
400	14.3	.0698	5.60	----	----	-----	42.5	49.7	.705	44.0	51.7	.730	46.5	53.5	.764	31 0	41.5	.686
25	28.6	.1396	1.075	72.0	73.7	1.018	71.5	73.2	1.039	63.5	67.7	.957*	56.2	68.5	.978*	52.0	64.0	1.057*
100	28.6	.1396	4.300	58 5	61.7	.852	54.2	59.7	.847	51.5	57.8	.816	50.5	55.25	.788	44.3	56.2	.929
225	28.6	.1396	9.68	48.5	54.8	.756	42.0	48.0	.681	36.5	46.9	.663	39.0	47.30	.675	31.0	40 2	664
400	28.6	.1396	17.20	39.0	45.2	.623	30.5	37.2	.528	27.5	41.7	.589	29.0	36.1	.515	21.5	29 7	491
25	57.2	.279	3.25	60.0	63.4	.875	49.0	56.7	.805	50 0	57.2	.808*	44.5	54.9	.782*	39.5	53.3	.881
100	57.2	.279	13.00	42.0	47.2	.651	31.5	38.8	.550	32.5	40.6	.573	30.3	40.7	.581	27.8	41.1	.680
225	57.2	.279	29.20	30.5	37.8	.522	23.0	29.7	.421	22.5	31.0	.438	20 0	27.0	.386	16.5	26.6	.440
400	57.2	.279	52.00	28.5	36.2	.500	19.0	25.5	.362	17.0	26.7	.377	19.5	20 27	.290	12.3	17.7	.293
25	114.3	.558	9.82	42.5	48.6	.670	33.5	38.7	.549	30.5	37.7	.535*	26.5	35.0	.500*	24.3	36.2	.595*
100	114.3	.558	39.30	24.0	33.2	.458	20.0	28.0	.397	17.0	25.7	.363	15.5	21.7	.310	11.0	18.0	.297
225	114.3	.558	88.40	18.0	25.2	.348	18.0	21.5	.305	15.0	20.5	.290	12.5	17.5	.250	9.0	13.6	.225
400	114.3	.558	157.20	17.5	24.7	.341	17.5	17.7	.251	14.5	19.5	.276	11.5	16.5	.236	8.5	14.2	.233
			φ_o		72.5			70.5			70.8			70.0			60.5	

the corresponding angular location of the sampling tubes as indicated on the abscissa. The spray angles referred to in our sampling data may be considered a convenient measure of fuel distribution along a circular arc 4$^1/_2$ in. from the nozzle tip.

Two possible sources of error were checked early in our tests: (*a*) At the high fuel and tank pressures considerable mist would form in the tank. By observation of the spray shape during sampling tests, it was determined that sampling error due to the presence of the mist was negligible. (*b*) Since the nozzle was spraying vertically downward, it was necessary to consider whether gravitational forces were affecting the spray distribution. That such was not the case was established by spraying a nozzle horizontally and noting that the spray remained symmetrical about the nozzle axis for at least 9 in.

The curves in Fig. 6 show the fuel distribution for a 60-gph 80-deg nozzle at 100-psi nozzle Δp and various tank pressures. For each test condition, a distribution curve was obtained for each of the ten nozzles. This gives an accurate representation of the fuel distribution at each condition of nozzle and tank pressure. In order to describe nozzle performance in terms of the variables involved, conveniently, it was necessary to reduce each curve to a single numerical value. Of the various methods for doing this, two have been selected.

The first method involved reducing the individual distribution curve to a "diametral spray angle"—ψ. This spray angle was calculated by finding the value of $\psi = \psi_L + \psi_R$ where ψ_L and ψ_R are equal to

$$\frac{\Sigma\theta\Delta\theta y}{\Sigma y\Delta\theta} = \frac{\Sigma\theta y}{\Sigma y}$$

for the left and right lobe of the distribution curve, respectively. As shown in Fig. 5, θ is the angular location of a sampling tube measured from the nozzle axis; $\Delta\theta$ is the angle between adjacent sampling tubes, i.e., 5 deg; y is the ordinate of the fuel-distribution curve.

The second method involves reducing the individual distribution curve to an "equivalent spray angle"—φ. This angle is calculated for each distribution curve by finding the value of $\varphi = \varphi_L + \varphi_R$ where φ_L and φ_R are equal to

$$\frac{\Sigma 2\pi r y\theta\Delta\theta' \sin\theta}{\Sigma 2\pi r y\Delta\theta' \sin\theta} = \frac{\Sigma y\theta \sin\theta}{\Sigma y \sin\theta}$$

for the left and right side of the distribution curve, respectively. The symbols here are the same as before with r being the radial distance from the nozzle tip to the sampling tubes and $\Delta\theta' = 2$ deg. The value of $\Delta\theta'$ was taken smaller than the previously used $\Delta\theta$ in order to give better accuracy. Calculations have shown that φ differs at most by 1 to 2 deg from the vertex angle of a cone which contains half of the total fuel spray at a 4$^1/_2$-in. radius from the nozzle. It is assumed here that the spray is symmetrical about the nozzle axis. This assumption was verified in several instances by showing that rotation of the nozzle about its axis had negligible effect on the spray distribution recorded. Values of φ versus nozzle pressure drop, with tank pressure as the parameter, are shown for the 60-gph 80-deg nozzle in Fig. 7, while values of φ versus tank pressure, with nozzle pressure drop as the parameter for the same nozzle, are shown in Fig. 8. In Table 2, the values of ψ, φ, and φ/φ_o are given for all the nozzles tested together with the ambient pressure and nozzle pressure drop at which they

TABLE 2 SPRAY-DISTRIBUTION DATA (*continued*)

TEST CONDITIONS				NOZZLE													
Nozzle Press. Drop P, psi	Tank psia	Gas Density γ lb/ft.^{3}	$P\gamma^{1.6}$	100 GPH - 45°			60 GPH - 45°			45 GPH - 45°			20 GPH - 45°			9 GPH - 45°		
				ψ	φ	φ/φ_0	ψ	φ	φ/φ_0	ψ	φ	φ/φ_0	ψ	φ	φ/φ_0	ψ	φ	φ/φ_0
25	1.5	.00732	.00963	44.6	47.0	1.290*	43.2	44.5	1.220*	43.0	44.1	1.085*	28.3	30.5	.820*	----	----	-----
100	1.5	.00732	.0385	44.5	44.8	1.229	44.0	45.0	1.234	45.5	46.1	1.132	40.8	43.0	1.156	38.8	42.8	1.26*
225	1.5	.00732	.0866	44.4	44.0	1.205	44.5	45.5	1.248	46.3	47.2	1.160	43.8	44.5	1.196	41.0	43.8	1.288
400	1.5	.00732	.1540	----	----	-----	44.5	45.4	1.245	46.0	46.5	1.141	44.3	44.5	1.196	40.0	43.6	1.281
25	3.58	.01748	.0375	43.2	45.7	1.253*	41.4	43.2	1.185*	40.5	42.0	1.032*	25.0	29.2	.785*	----	----	-----
100	3.58	.01748	.150	42.8	43.3	1.187	41.4	43.2	1.185	43.5	45.2	1.110	38.8	41.8	1.122	37.4	40.7	1.197*
400	3.58	.01748	.600	----	----	-----	38.8	40.8	1.120	41.4	44.0	1.080	39.7	42.4	1.140	37.2	40.3	1.185
25	7.15	.0349	.1176	40.8	43.2	1.183*	38.8	41.4	1.134*	37.5	39.4	.968*	21.7	27.2	.731*	----	----	-----
100	7.15	.0349	.470	40.2	41.0	1.123	38.4	40.7	1.117	40.7	44.0	1.080	34.0	40.0	1.075	34.7	37.8	1.110*
400	7.15	.0349	1.880	----	----	-----	33.3	35.7	.978	35.7	39.4	.968	36.8	38.3	1.030	32.7	36.3	1.068
25	14.3	.0698	.350	37.5	39.6	1.086*	34.9	39.0	1.070*	34.0	36.7	.902*	18.0	24.0	.645*	----	----	-----
100	14.3	.0698	1.40	34.8	36.5	1.000	33.8	36.5	1.000	36.2	40.7	1.000	34.5	37.2	1.000	30.3	34.0	1.000*
225	14.3	.0698	3.15	33.2	35.0	.960	30.8	32.8	.899	31.4	36.6	.899	33.3	33.8	.909	30.0	34.0	1.000
400	14.3	0698	5.60	---	---	- --	25.2	28.8	.790	27.7	30.6	.752	27.3	31.2	.839	25.4	30.4	.895
25	28.6	.1396	1.075	34.2	35.7	.978*	32.0	36.0	.987*	31.7	35.0	.860*	19.8	26.0	.698*	----	----	-----
100	28.6	.1396	4.30	25.5	28.5	.782	26.7	30.3	.831	29.2	32.0	.786	30.0	33.0	.888	23.8	29.2	.858*
225	28.6	1396	9.68	21.4	23.3	.638	20.0	24.0	.658	23.0	23.0	.565	21.4	24.8	.667	21.3	26.9	.791
400	28.6	.1396	17.20	18.5	21.5	.589	16.0	18.5	.507	17.0	19.5	.479	16.0	20.8	.559	16.1	21.7	.638
25	57.2	.279	3.25	25.2	27.5	.754*	24.3	26.8	.735*	24.5	27.7	.681*	19.3	23.8	.640*	----	----	-----
100	57.2	.279	13.00	16.8	20.6	.565	16.2	19.2	.527	17.0	21.0	.516	16.8	20.2	.543	13.2	16.7	.491*
225	57.2	.279	29.20	14.5	17.0	.466	11.3	16.5	.452	11.7	15.3	.376	12.2	17.2	.462	9.7	13.7	.402
400	57.2	.279	52.00	11.3	15.5	.425	10.5	16.2	.444	9.5	13.8	.339	9.5	13.7	.368	7.0	10.0	.294
25	114.3	.558	9.82	14.2	17.4	.477*	14.0	16.8	.461*	14.0	17.0	.417*	14.0	16.5	.445*	6.0	9.8	.288*
100	114.3	.558	39.3	10.0	15.4	.422	8.8	13.0	.357	10.3	15.6	.383	8.8	13.1	.352	7.0	10.8	.318*
225	114.3	.558	88.4	9.2	13.2	.362	7.7	11.5	.315	9.5	13.3	.327	8.0	13.1	.352	6.5	10.3	.303
400	114.3	.558	157.2	7.8	11.5	.315	7.5	11.0	.301	8.5	12.5	.307	8.4	13.0	.350	7.3	8.1	.238
			φ_0		36.5			36.5			40.7			37.2			34.0	

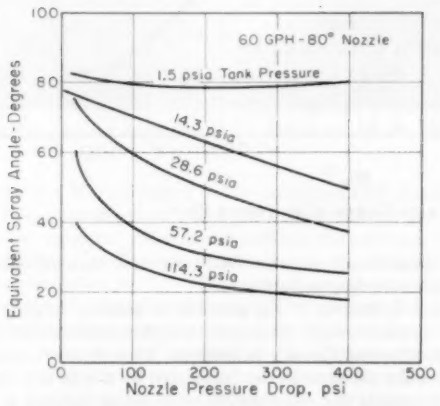


FIG. 7 EQUIVALENT SPRAY ANGLE VERSUS NOZZLE PRESSURE DROP

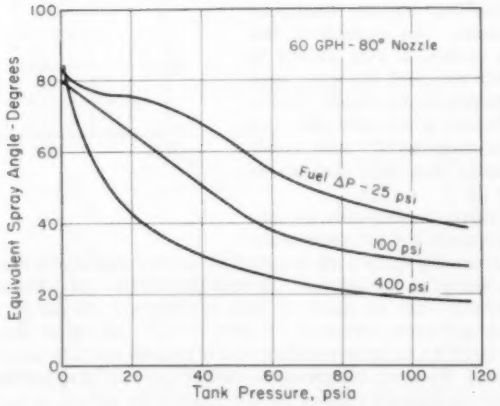


FIG. 8 EQUIVALENT SPRAY ANGLE VERSUS TANK PRESSURE

were obtained. The φ_0-value is the equivalent spray angle at a tank pressure of 14.3 psia and nozzle pressure drop of 100 psi. The values at tank pressures of 3.58 and 7.15 psia were obtained by interpolation from the ψ or φ versus tank-pressure curves. Blank spaces represent points where data could not be obtained because of the formation of a fuel bubble or Weber-number effects, except for 100-gph nozzles where blanks represent points at which fuel flow was limited due to cavitation at the fuel pump.

It was found that φ/φ_0, for all the nozzles tested, correlates well with $P\gamma^{1.6}$, where P = nozzle pressure drop in psi and γ = ambient-gas density in lb per cu ft. Points for which $\varphi_0 \mathrm{We} < 8 \times 10^4$ are distinguished in Table 2 by asterisks and are excluded from the plots of φ/φ_0 versus $P\gamma^{1.6}$. φ_0 is in degrees and the Weber number, $\mathrm{We} = Pd/\sigma$ with P = nozzle pressure drop, d = nozzle orifice diameter, and σ = surface tension of fuel. For these points, the bubble or Weber-number effect was predominant in the spray. To bring these points into good alignment on the curves would require considerable complication of the expression

$P\gamma^{1.6}$. It was therefore considered desirable to disregard this small number of points and retain the simple expression. Fig. 9 is a plot of all values of φ/φ_0 obtained except those excluded by the $\varphi_0 We < 8 \times 10^5$ criterion. Fig. 10 represents a curve drawn through the average values of the φ/φ_0 points shown in Fig. 9. From these figures, it can be seen that the change in φ/φ_0 is small for $P\gamma^{1.6} < 1$, becomes larger for $1 < P\gamma^{1.6} < 60$, and begins to decrease for $P\gamma^{1.6} > 60$.

One explanation of the decrease of spray angle with increasing P which has been advanced (8) assumes that increasing the air density at the core of the fuel vortex increases the frictional losses there. This, in turn, increases the ratio of axial to tangential velocity, thus producing a spray-angle decrease. We believe that this effect is not an important one for the following reasons: (*a*) A nozzle constructed with a solid core at the fuel vortex (i.e., no air core) still produced spray-angle reduction with increasing P and γ. (*b*) Photographs of the spray while showing a curvature in the spray outline indicate that the spray-cone angle, as defined in the Introduction, remains essentially constant. An example of this can be seen in Figs. 11 and 12 which represent the spray from a 60-gph 80-deg nozzle at tank pressures of 1.5 psia and 114.3 psia, respectively, with nozzle pressure drop held constant at 100 psi.

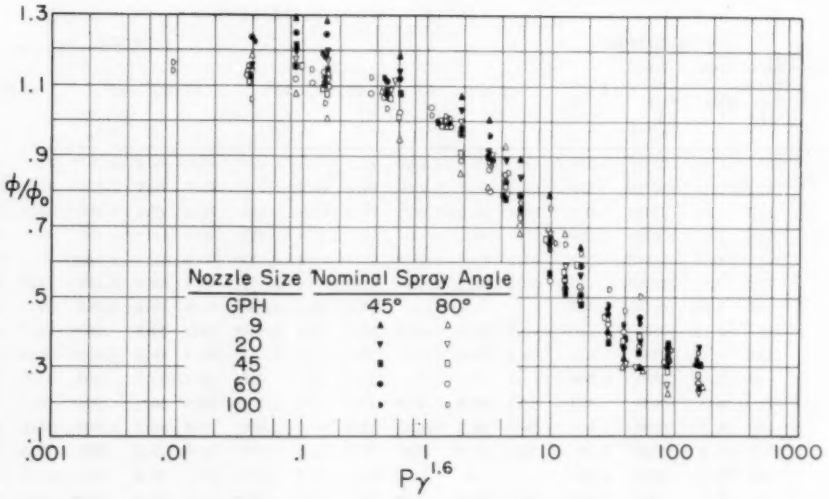


FIG. 9 INDIVIDUAL DATA POINTS, φ/φ_0 VERSUS $P\gamma^{1.6}$

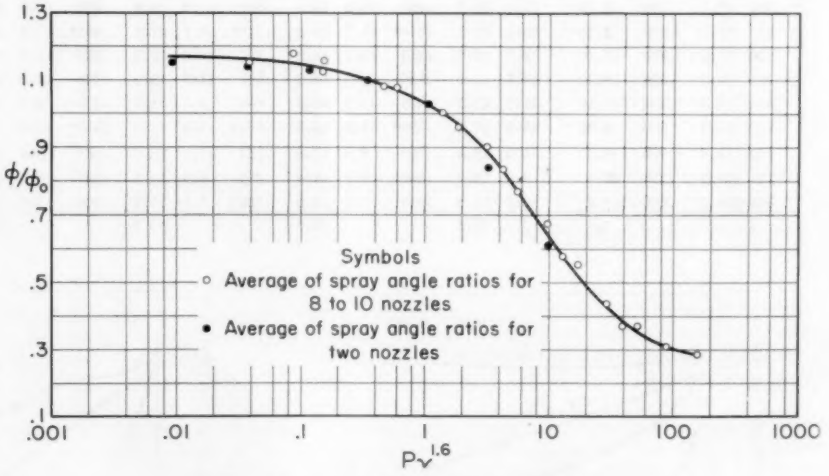


FIG. 10 AVERAGE CURVE, φ/φ_0 VERSUS $P\gamma^{1.6}$

From our experiments, the indications are that the phenomenon of decreasing spray angle is caused by aerodynamic effects due to the motion of the fuel spray through the ambient gas. The fuel emerging from the nozzle at high velocity entrains gas at the inner and outer surfaces of the spray sheath. (By spray sheath we mean an imaginary surface which encloses the fuel spray in space.) However, the gas supply to the inner portion of the spray sheath is limited by the enclosed volume in the sheath, as shown in Fig. 16. Of course for $\varphi = 180$ deg, the gas-flow path would be equally unrestricted for both sides of the spray sheath; but, we are dealing here with the more common type of nozzles where φ is generally less than 90 deg. Pressure differences resulting from this effect ($P_3 - P_1$ in Fig. 16) have been measured previously by one of the authors for swirling annular air jets used on air-assisted fuel nozzles (9). Pressure-difference data were obtained in the current tests for a 45-gph 80-deg nozzle and are shown plotted in Figs. 13, 14, and 15. Figs. 13 and 14 are plots of spray-sheath pressure difference versus tank pressure and versus nozzle pressure drop, respectively. Fig. 15 presents equivalent spray-angle versus spray-sheath pressure difference with tank pressure and nozzle pressure drop as parameters.

In some instances, it was possible to observe droplets being carried upward toward the nozzle along the nozzle axis by the reverse gas flow established. In addition, it was found that supplying gas under pressure at the nozzle axis by means of a tube inserted up inside the spray sheath produced an increase in spray angle. Thus we see that a low-pressure region exists inside the spray sheath and it appears that this is instrumental in causing the spray-angle decrease. The mechanism by which this pressure difference causes a decrease in spray angle is proposed to be as follows.

Where the spray sheath exists as a liquid sheet, the pressure difference acts directly on this sheet. The distance from the nozzle orifice for which this sheet persists may be very small at the higher values of P and γ. In the region where the fuel spray consists of droplets, Fig. 16 shows the static-pressure condition existing in a plane perpendicular to the nozzle axis, with $P_2 < P_1 < P_3$ where

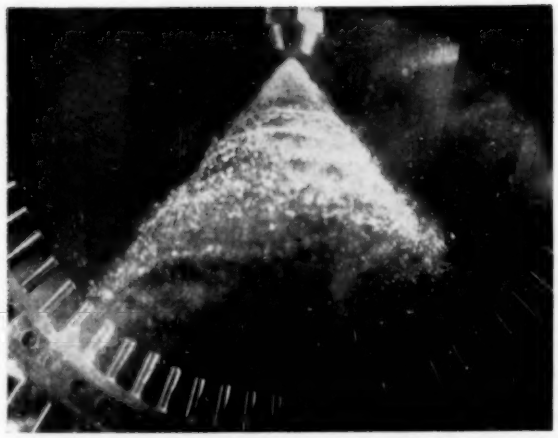


FIG. 11 FUEL SPRAY AT LOW TANK PRESSURE (60/80 DEG AT 1.5 PSIA AND 100 PSI ΔP)

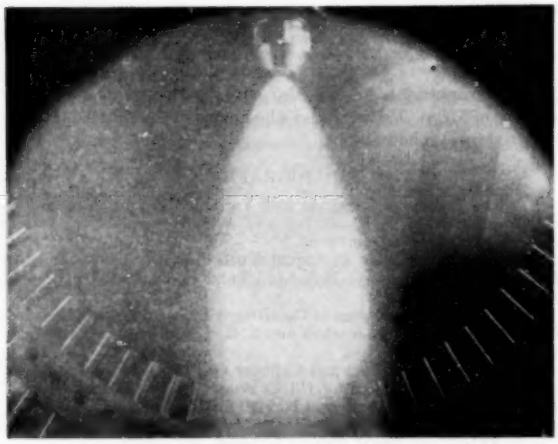


FIG. 12 FUEL SPRAY AT HIGH TANK PRESSURE (60/80 DEG AT 114.3 PSIA AND 100 PSI ΔP)

P_2 represents a mean pressure in the combined fuel droplet-air stream, P_1 is a mean pressure inside the spray sheath, and P_3 is the ambient pressure. The arrows indicate the probable directions of gas flow. In Figs. 13, 14, and 15, the spray-sheath pressure difference refers to $P_3 - P_1$, where P_1 is measured on the nozzle axis, $^1/_2$ in. from the nozzle tip. The pressure P_2 was not actually measured but must be taken to be less than P_1 and P_3 if gas is to be entrained from regions 1 and 3. Thus, in the region where the fuel breaks up into droplets, the spray sheath may be considered to be droplets and entrained gas. The pressure difference $P_3 - P_1$ sets up air flows which produce droplet acceleration toward the nozzle axis.

According to the preceding analysis, if a nozzle were constructed having a spray-cone angle of 180 deg, there should be no reduction in spray angle since the volumes inside and outside the spray sheath are equal. Thus one would expect the effect of P and γ on spray angle to decrease as the spray-cone angle increases. One might ask, then, how the φ/φ_0 versus $P\gamma^{1.5}$ curves for both the 45 and 80-deg nozzles fall so close together. This may be visualized as the result of two opposing effects. While the pressure drop generated across the spray sheath increases with decreasing cone angle, the space available for a change in φ is

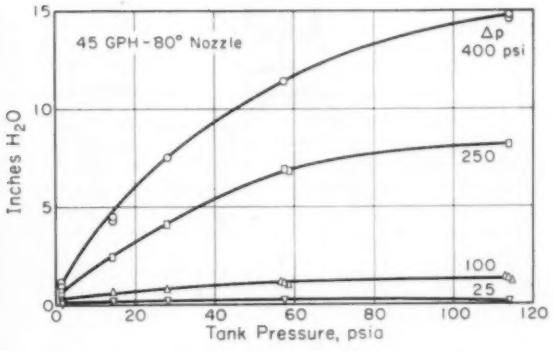


FIG. 13 SPRAY-SHEATH PRESSURE DIFFERENCE VERSUS TANK PRESSURE

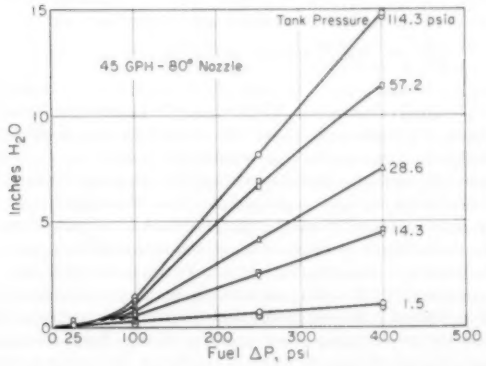


FIG. 14 SPRAY-SHEATH PRESSURE DIFFERENCE VERSUS FUEL PRESSURE ΔP

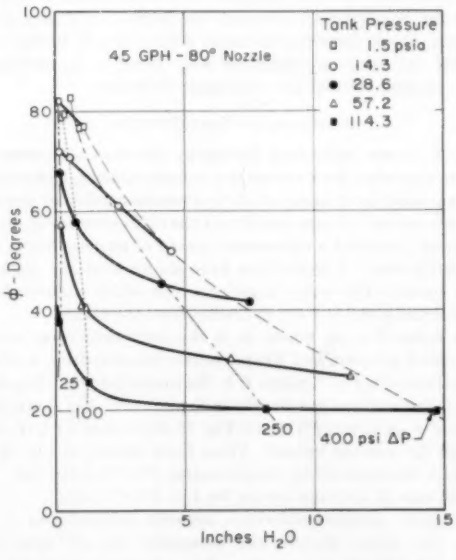


FIG. 15 EQUIVALENT SPRAY ANGLE VERSUS SPRAY-SHEATH PRESSURE DIFFERENCE

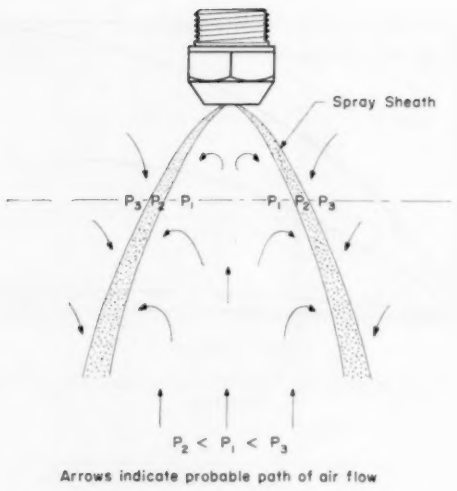


FIG. 16 SPRAY-SHEATH PRESSURE DIFFERENCES

correspondingly decreased. Thus, when the results are plotted in the form of a ratio φ/φ_0, these effects tend to cancel giving essentially one curve on the φ/φ_0 versus $P\gamma^{1.6}$ plot.

From photographs and observations of the spray through the windows in the tank, it was concluded that the drop size of the spray decreased with increasing tank pressure and nozzle pressure drop. An example of the difference in spray-droplet appearance can be seen by comparing Figs. 11 and 12 where the tank pressure increased from 1.5 to 114.3 psia with the nozzle pressure drop constant at 100 psi. No actual drop-size measurements were made.

Since the aerodynamic effects are so pronounced and are related to the pressure difference which is produced between the inside and outside of the spray sheath, it was interesting to find the effect of enclosing the fuel spray with a combustor shape. This was done using a typical combustor shape consisting of a hemisphere followed by two cylindrical sections all of 8 in. diam. Holes were put in the hemisphere to simulate the normal open area at that location. Spray distributions taken with a 60-gph 80-deg nozzle over the full range of conditions were found to be identical to those obtained without the combustor enclosure.

SUMMARY OF TEST RESULTS

Fig. 6 shows individual fuel-spray distribution curves. In order to represent these curves in a concise manner, a characteristic angle must be defined which is a representation of these distribution curves. It was found that the equivalent spray angle φ, as defined, provided a convenient means of representing a given fuel distribution. Calculations have shown that the angle φ is nearly equal to the vertex angle of a cone which contains half of the total fuel spray at a $4^1/_2$-in. radius from the nozzle.

The ratio of φ/φ_0, where φ_0 is the equivalent spray angle at atmospheric pressure and 100-psi nozzle pressure drop, was found to correlate with $P\gamma^{1.6}$, where P is the nozzle pressure drop in psi, and γ is the ambient gas density in lb per cu ft. Fig. 9 is a plot of values of φ/φ_0 versus $P\gamma^{1.6}$ and Fig. 10 represents a curve drawn through the average values. From these figures, it can be seen that φ/φ_0 decreases when the parameter $P\gamma^{1.6}$ is increased. The greatest rate of decrease occurs for $1 < P\gamma^{1.6} < 60$.

The static pressure difference between ambient and a point inside the spray sheath was measured for all test points with a 45-gph 80-deg nozzle. For all cases, it was found that the pressure inside the spray sheath was lower than the ambient pressure. This pressure difference increased with increasing tank pressure and nozzle pressure drop and appears to be instrumental in causing the spray-angle decrease. An explanation of the spray-angle decrease is given based on the aerodynamic effects caused by motion of the fuel spray in the ambient gas.

Spray distributions taken with the nozzle surrounded by a typical combustor upstream section were found to be identical with those taken without the combustor enclosure.

Of course, the spray distributions as obtained here will be modified by the combustor air flow during actual burning; to what extent is not known precisely. From past experience in combustion tests, we know that a change from an 80 to 60-deg or 60 to 45-deg nozzle results in considerable variations in combustion efficiency, flame length, and combustor-wall temperatures. From this, we may surmise that the equivalent spray angle is an important factor and that the variation in equivalent spray angle must be considered in cases of combustor scaling or new designs where differences in fuel or ambient gas pressures will be encountered. An example of such a case would occur in attempting to predict the performance of a combustor at operating pressures of the order of 6 atm from combustion tests at atmospheric pressure.

ACKNOWLEDGMENTS

The guidance of Dr. A. E. Hershey during the course of this project is appreciated. We also wish to acknowledge the assistance of Mr. A. T. Pieczynski who helped in compiling data and took the spray photographs.

BIBLIOGRAPHY

1 "The Design of Constant and Variable-Capacity Mechanical Oil Atomizers," by J. F. Harvey and A. S. Hermandorfer, Trans. SNAME, vol. 51, 1943, pp. 61–82.

2 "The Atomization of Liquid Fuels," by E. Giffen and A. Muraszew, Chapman and Hall Ltd., London, England, chapter 4, 1953.

3 "Factors in the Design of Centrifugal Type Injection Valves for Oil Engines," by W. F. Joachim and E. G. Beardsley, NACA Report No. 268, 1927.

4 "The Effects of Fuel and Cylinder Gas Densities on the Characteristics of Fuel Sprays for Oil Engines," by W. F. Joachim and E. G. Beardsley, NACA Report No. 281, 1927.

5 "Oil-Spray Investigations of the N.A.C.A.," by W. F. Joachim, Trans. ASME, vol. 49–50, Paper OGP-50-6, 1927–1928.

6 "A Comparison of Fuel Sprays From Several Types of Injection Nozzles," by D. W. Lee, NACA Report No. 520, 1935.

7 "Combustion and Combustion Equipment for Aero Gas Turbines," by E. A. Watson and J. S. Clarke, *Journal of the Institute of Fuel*, vol. 21, no. 116, Oct., 1947, pp. 1–34.

8 "The Atomization of Liquid Fuels for Combustion," by J. R. Joyce, *Journal of the Institute of Fuel*, vol. 22, 1949, pp. 150–156.

9 "Air-Assisted-Fuel Nozzle Development," by S. M. DeCorso, Westinghouse Research Laboratories, Pittsburgh, Pa., Research Report R-94451-1-C, 1954.

Discussion

H. CLARE.[3] A most interesting feature of this paper is the novel method of assessing spray-cone angle, which is shown to vary with applied fuel pressure even though a silhouette of the spray cone would appear to be sensibly constant. The phenomenon of the change in cone angle with ambient pressure is also well illustrated. However, the hypothesis would be more conclusive if results were included to show the effect of ambient pressure on the flow number of the nozzle. If the effect of high-density gas at the air core of the nozzle is to increase friction losses in the fuel vortex, and hence the ratio of the axial and tangential velocities, it is reasonable to expect that any reduction in

[3] Chemical Physics Dept., Ministry of Supply, National Gas Turbine Establishment, Pyestock, Farnborough, Hants, England.

spray-cone angle due to this mechanism would be accompanied by an increase in the discharge coefficient.

Authors' Closure

While we are in general agreement with Mr. Clare's remarks, we note that the definition of the term "spray-cone angle" as used in his discussion evidently differs from that given in the introduction of the paper. Thus, we would say that the spray-cone angle remained essentially constant while the spray angle (as measured at the sampling tubes) varied. Concerning the flow-number variation, an interesting point is raised which might well have been covered in the paper. As Mr. Clare points out, any increase in frictional losses at the fuel vortex should cause an increase in the flow number. Examination of the flow-rate data (not given in the paper) showed no increase in flow number with increasing ambient pressure, indicating that the increase of frictional loss at the fuel vortex is negligible.

The following additional test results were presented orally by Mr. DeCorso at the Gas Turbine Power Conference because it was felt that the scope of application of the data was increased through this additional information. We have, therefore, included these additional results in the discussion.

Fig. 10 of the paper gives the relation between φ/φ_0, the dimensionless equivalent spray angle, and $P\gamma^{1.6}$. Data for this curve were obtained from ten nozzles of different capacities but all having spray-cone angles of either 45 or 80 deg. These nozzles were all manufactured by the same company. To increase the usefulness of this curve, it obviously would be desirable to determine whether it applied for nozzles made by other manufacturers and nozzles of different spray-cone angles. Fig. 17 of this closure is a plot of this same curve with the additional data points representing nozzles of different manufacturers and additional spray angles of 30, 60, and 100 deg. It can be seen that only the points for the 60/100-deg nozzle are significantly displaced from the curve. As the spray-cone angle increases toward 180 deg, there

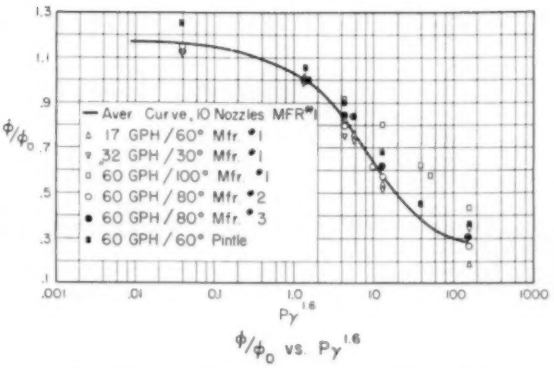


Fig. 17 Average Curve and Additional Data

should be less reduction in spray angle with changes in pressure and density because the aerodynamic effects inside and outside the spray sheath become more and more similar. The 60/60-deg pintle nozzle was manufactured in our shop. This nozzle without any fuel swirl showed spray-angle variations similar to those encountered with the other nozzles. These results indicate that the curve predicts the spray-angle changes for nozzles having listed spray-cone angles varying from 30 to 80 deg. This is the range of spray-cone angles generally encountered in fuel-oil-burning gas turbines and many other combustion installations. From the observation that nozzles by different manufacturers showed the same variation of spray angle with changes in nozzle pressure and ambient gas density, it can only be concluded again that these spray-angle changes are primarily produced by aerodynamic effects acting on the spray after it has left the nozzle tip and that variations in the internal geometry of the nozzle do not affect results significantly.

An Experimental Arrangement for the Measurement of the Pressure Distribution on High-Speed Rotating Blade Rows

By K. LEIST,[1] AACHEN, GERMANY

For several years past the research staff of the Institute for Turbomachines of the Aachen Technical University has carried out measurements on rotating turbine blading. This program is part of a comprehensive effort directed toward the experimental investigation of the three-dimensional flow through axial-flow turbomachines.

Introduction

MEASUREMENTS on rotating blades are of interest because the effect of rotation of the row, and of the centrifugal forces arriving therefrom, upon the flow of the working medium so far has been explored but slightly. It scarcely can be expected that this very complicated flow can be determined precisely in its entirety. However, it is intended to gather a maximum of significant information from measurements on coaxial cylindrical sections. This information may not be adequate as a basis for the establishment of a generally valid method of analytical treatment of the flow because of the elusive nature of the mutual influence of numerous contributing influences. Nevertheless, the results of such measurements should be an important contribution to the three-dimensional theory of axial-flow turbomachines.

Experimental Setup

To start with we shall discuss the experimental arrangement. It was developed for the purpose of measuring the pressure distribution over the surface of the blades of rotating axial or radial-flow-turbine or compressor-blade rows. The development and construction of this experimental arrangement was undertaken and carried to completion at the Aachen Technical University.

Some mechanisms previously described in the literature have been used by, e.g., Fuhrmann, Betz and Mautz, Keller and Bleuler, Weske, Runckel and Davey, Himmelskamp, Morelli and Bowerman. Apart from the fact that all except Keller's investigation pertain to retarding flow, that is, compressor blading, the speeds of the previous devices are about 2000 rpm and less; and the peripheral speeds exceed the 100 m/sec limit only in the case of the Runckel and Davey investigation which is 117 m/sec.

In the arrangement at Aachen University, peripheral velocities of 170 m/sec are produced with speeds of about 10,000 rpm. Consequently the diameters of the rotor are so small that the centrifugal accelerations, the effect of which upon pressure distribution is one of the chief objectives of our investigation, are increased sevenfold relative to the others. So we hope that the influence of the centrifugal force upon the flow will be rendered as precisely as possible. Measurements in the wake of rotor blades by means of a rotating probe were made by Weske who reported important results.

The first version of the test setup at Aachen University, developed by W. Fister,[2] which already had solved some questions of the problem, required several important improvements, some of them of a practical nature. It was found necessary, for instance, to reduce wear of instrumental components to insure reliable operation over extended periods.

On the basis of this experience an improved experimental arrangement was developed by W. Dettmering. This second setup has given satisfactory service for about 250 hr of operation with exactly reproducible results at various operating conditions and speeds. It was used for a comprehensive program of systematic measurements.

The principal differences to be expected between the flow past fixed straight grids and the flow in normal operation of a turbomachine past rotating blades are as follows:

(*a*) Effect of variation of pitch along the radius because of the fanning of the blades in the rotor wheel. This effect can be determined from investigations of the fixed cylindrical grid.

(*b*) Centrifugal effects in the boundary layer along the wall of the blades and the radial secondary flow arising from it.

(*c*) Variation of pressure along the radius ahead of the rotor blades and the radial-flow component arising therefrom, especially for long blades.

(*d*) Tip leakage through the clearance between rotor blade and casing and its influence upon pressure distribution near the tips of the blades.

(*e*) Wake flow behind the stator blades.

(*f*) Variation, along the length of the blades, of the direction of flow at the leading edges.

The second experimental arrangement for the measurement on rotating blading is shown in Fig. 1. The turbine is coupled to a hydraulic torque dynamometer to determine its output. Fig. 2 shows the over-all arrangement of the entire unit in a test cell. The cold-air flow through the turbine is induced by suction by means of a compressor. Much care has been taken to achieve satisfactory axially symmetrical inflow at the stator wheel; all disturbing influences have been eliminated. The stator row may be adjusted circumferentially with center in the shaft axis by a manually operated mechanism. The essential parts of the instrumental equipment are described with reference to Fig. 3 as follows:

[1] Professor, Technical University.

Presented at the Gas Turbine Power Division Conference, Washington, D. C., April 16–18, 1956, of The American Society of Mechanical Engineers.

Note: Statements and opinions advanced in papers are to be understood as individual expressions of their authors and not those of the Society. Manuscript received at ASME Headquarters, March 15, 1956. Paper No. 56—GTP-13.

[2] "Druckverteilungsmessungen an umlaufenden Turbinenschaufeln," by W. Fister, "VDI-Forschungsheft" 448, 1955.

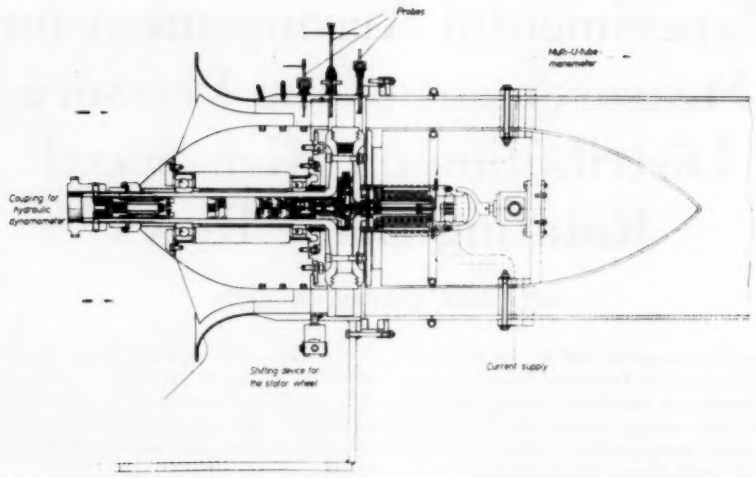


FIG. 1 TURBINE TEST STAND FOR MEASUREMENT OF PRESSURE DISTRIBUTION ON ROTATING BLADES. MAXIMUM SPEED n = 13,000 RPM

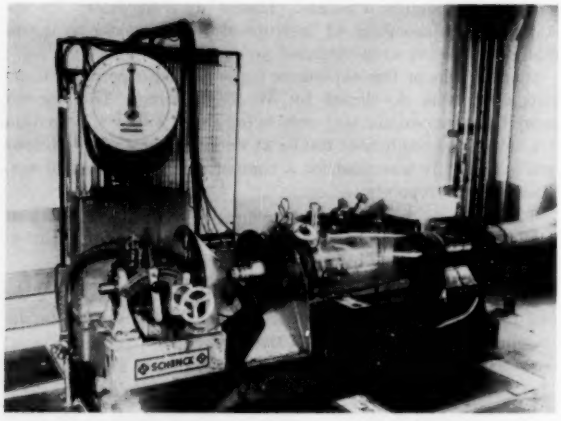


FIG. 2 OVER-ALL VIEW OF EXPERIMENTAL UNIT IN TEST CHAMBER

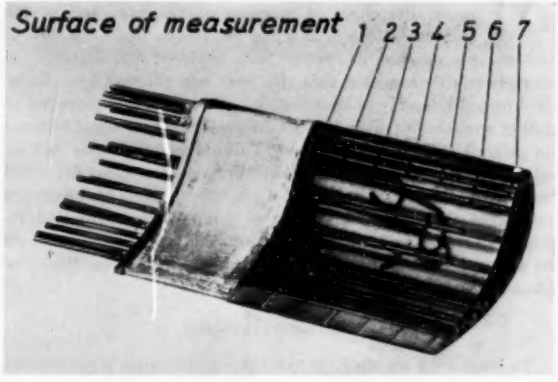


FIG. 4 PLASTIC BLADE WITH PRESSURE HOLES IN SEVEN PLANES OF MEASUREMENT

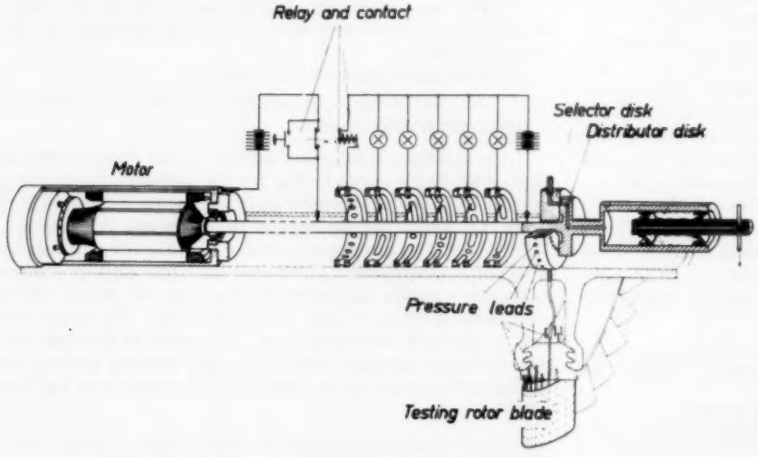


FIG. 3 SCHEMATIC DRAWING OF MEASURING DEVICE

The testing rotor blade made of a fusible plastic (Araldit Giessharz B) and glued into an aluminum ring was provided with 20 cast-in pressure conduits of 1 mm diam. Connections to each of the conduits of 0.4 mm diam were drilled normal to the surface of the blade. This blade was tested up to the breaking speed of 17,000 rpm. There are 20 such holes distributed over the front and back of the profile in each surface of measurement, Fig. 4. Altogether seven surfaces of measurement are provided at various radii from tip to root of the blade; that is, 140 holes on one blade. The locations of the measuring points are shown in Fig. 5. The pressure distribution in only one of these surfaces can be measured in any one test run, the pressure holes of the remaining six being closed off with thin adhesive plastic tape (Fig. 7).

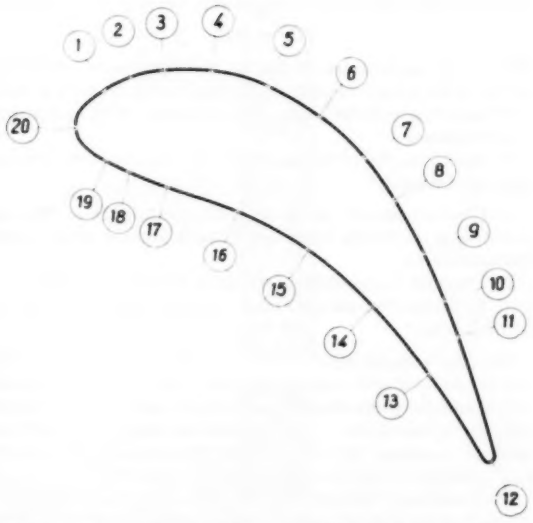


FIG. 5 DISTRIBUTION OF PRESSURE HOLES AROUND BLADE PROFILE

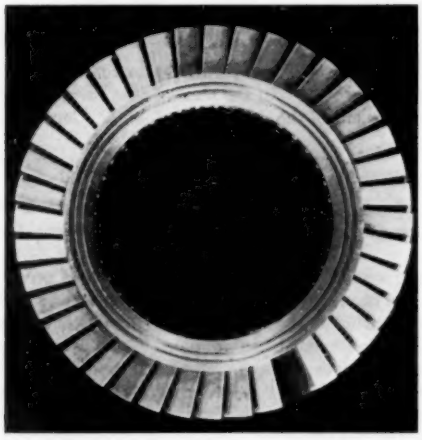


FIG. 6 WHEEL ASSEMBLY WITH BLADE FOR MEASUREMENTS

A test wheel is shown in Fig. 6. The 20 pressure leads are connected to terminals on a disk rotating with the wheel and then transmitted by a selector mechanism to a disk rotatable relative to the wheel. This disk is operated by an electric motor placed within the turbine rotor. The connecting sliding surfaces of the selector mechanism are designed carefully to prevent any leakage. In this way the pressures are transmitted without error from the rotating turbine shaft to the multi-U-tube manometer arranged outside. The rotating pressure seal, connecting the rotating and fixed parts of the leads, is kept to minimum diameter and sealed by two packing rings with intermediate oil chamber. Thereby friction and heat generation are kept to a minimum. Provision is made also for water cooling, yet this was found necessary only for the longest periods of measurement.

This type of seal between the rotating and fixed parts has given excellent service in many hours of operation. Pressures transmitted by the seal are successively applied to one of the 20 U-tubes by a distributor and selector disk of similar design to that

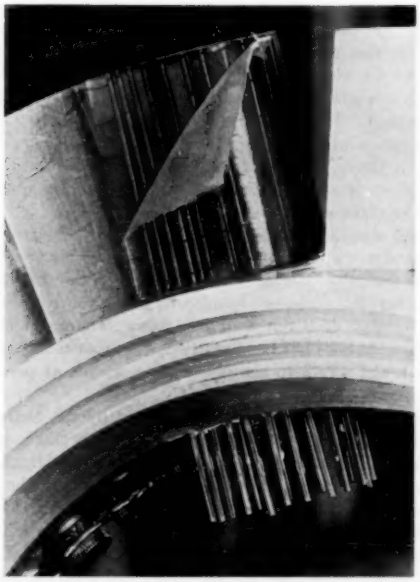


FIG. 7 VIEW OF BLADE WITH PLASTIC FILM COVERING PRESSURE HOLES PARTLY REMOVED

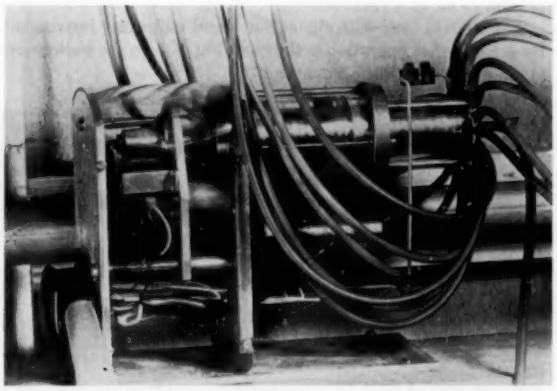


FIG. 8 DRIVE MOTOR FOR DISTRIBUTOR MECHANISM

in the turbine. In this way a complete representation of the pressure distribution is produced which then may be photographed. A further merit of this method lies in the fact that it provides an automatic check against leakage, as the pressure reading at the time of its measurement may be compared with the pressure read after a considerable time, e.g., one night, has elapsed. This check has been operated frequently.

OPERATION

As shown in Fig. 8, a second selector disk operated by a positively synchronous motor serves to distribute the pressures successively among the various pressure leads.

It is important that the two selector disks be synchronized in order that a particular pressure hole be co-ordinated with the corresponding manometer column. Available synchronous motors could not be selected for this purpose as their torque was insufficient. Instead, two d-c motors operated by relay control were chosen for this drive. The selector disks are rotated slowly by them through a reduction-gear train of 1:5000 gear ratio.

The operation of this mechanism is checked by a position indicator. For test stand II, an arrangement of electric bulbs was used indicating the position of the two selector disks. Current supply to the electric motor, initially by way of brushes and rings, now is admitted to the rotating system through standard ball bearings. This is a very satisfactory solution. The transition resistance which, in the case of brushes, rises sharply at high speeds, in the case of ball bearings remains essentially constant independent of speed. Consequently the voltage once chosen may be kept the same for all speeds.

As both selector disks, by the action of their respective motors, are rotated to the next bore of the distributor, a silver contact pin fixed on the selector disk and rotating with the disk touches a contact bar and, closing the circuit, stops the rotation of the motor by way of a relay. Each motor is stopped separately after the next bore has been reached. An electrically controlled friction brake serves to arrest the selector disk at precisely the desired position.

After a certain time interval controlled by the charging of a condenser, the length of which is chosen such that equilibrium pressure of the manometer liquid column is attained, the motors receive a new impulse and the next cycle begins. The duration of this impulse is freely adjustable and it must suffice to advance the selector disks beyond the pressure bores and to separate the contact pin and bar. Yet it must not be so great that the next bore is passed before the breaker relay has a chance to get into action. This completely automatic and widely adjustable selector mechanism for the 20 pressure taps of a profile section has proved to be very satisfactory.

The control desk with signal bulbs and indicating instruments for selector-disk operation is shown in Fig. 9. In the background is the test cell with the test stand.

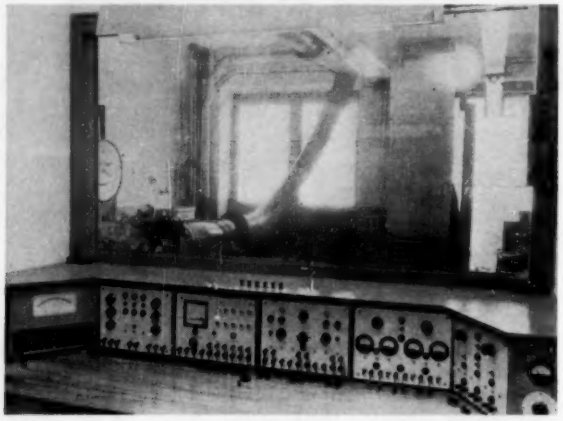


FIG. 9 VIEW OF CONTROL PANELS IN OPERATING STAND

The details of design of the pressure distributor arranged in the hollow shaft of the turbine rotor can be seen in Figs. 10 and 10(*a*). The apparatus is relatively extensive and complicated which is a necessary outcome because of the requirements of short periods of readings for the systematic measurement of sequences of indications, of high reliability of operation, and great accuracy. All this has been achieved successfully as can be shown by exact reproducibility of any reading. Simultaneous with the pressure-distribution measurements, readings are taken by means of flow-measuring probes of the inflow and outflow conditions for the rotating blading.

The effect of centrifugal forces upon the fluid in the pressure leads was taken care of by a correction factor K defined as follows

$$P_r = P_0 K$$

where P_0 is the pressure reading on the manometer board connected to the pressure lead from the seal on the center line shaft, P_r is the pressure at the point of measurement, both with respect to atmosphere.

To determine the correction factor K, one of several assumptions can be made

1 Constant density of the air in the rotating tube. This assumption is frequently made but it leads in some cases to considerable errors.

2 Constant temperature of the air in the rotating tube.

3 Polytropic temperature and pressure distribution in the rotating tube.

Since the assumption of constancy of air temperature throughout the rotating tube appears to come close to actual conditions it was used for computing K, particularly since it can be shown that slight variations of the temperature have little influence upon K; a change of 10 deg of the temperature assumed constant caused a 0.5 per cent change in the value of K.

By isothermal relations the correction factor is a function of the radii of the point of measurement r, and of the point of transition from the rotating to the stationary lead r_0, the speed of rotation $n = 60\ \omega/2\pi$, and the temperature T

$$K(r, n, T) = \frac{P_r}{P_0} = e^{\frac{\omega^2(r^2 - r_0^2)}{2gH_0}}$$

where $H_0 = RT_0 = P_0/\gamma_0$ pertaining to conditions in the fixed pressure leads $r = 0$. In the present case the pressure seal was on the center line of the shaft, hence $r_0 = 0$. Consequently

$$K(r, n, T) = \frac{P_r}{P_0} = \exp \frac{\omega^2 r^2}{2g\,RT_0}$$

The influence of variations of barometric pressure is eliminated by reducing the corrected pressures to nondimensional form $(P_0 - P_1)/q_1$.

For purposes of comparison, the same curve of pressure distribution has been shown in Fig. 11, first without any correction, next as corrected on the assumption (1) of constant density (γ = const), and finally on the assumption (2) of constant temperature T = const. An experimental check was made of the validity of the correction factor K by rotating a pilot tube which measured the total pressure of the peripheral velocity. The reading when corrected by the foregoing factor was in close agreement with the total pressure $P_{\text{stat.}} + \rho \cdot \frac{u^2}{2}$ at the point of measurement. A further check was made by filling the leads with different gases, e.g., air and hydrogen. The different readings, when corrected, resulted in the same value for the pressure at the point of measurement.

CHECK TESTS MADE

Since for the novel experimental arrangement under discussion a check was to be made through comparative measurements on the fixed grid, a preliminary investigation already had been made on stand I in which systematically all effects causing discrepancies between the fixed and moving grid had been eliminated. This was done by: choice of identical profiles, axial inflow through a straightener section with very thin walls (without inlet guide vanes); measurements on the pitch circle at very low rotative speeds; equal pitch in the fixed grid and on the pitch circle of the rotor grid.

Discrepancies between the two methods of investigation in this way were largely eliminated, namely:

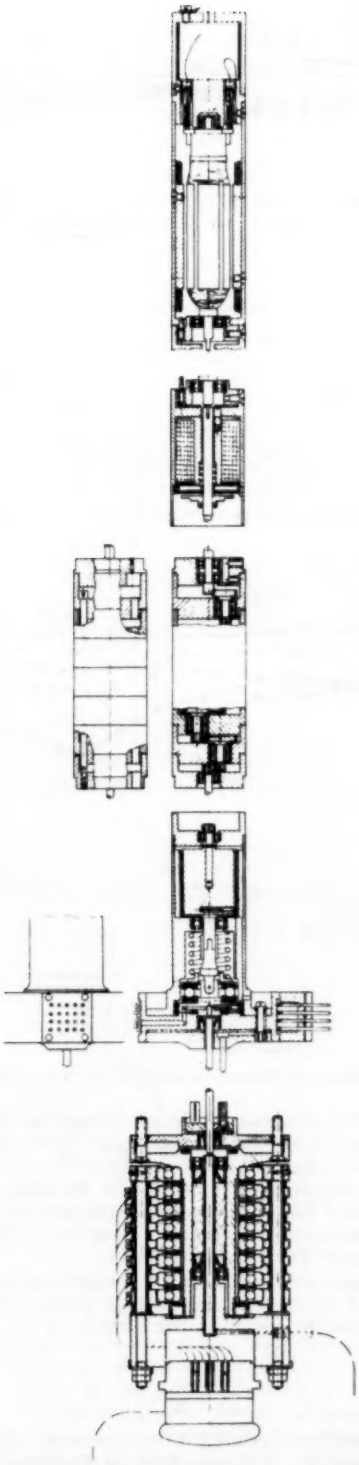


FIG. 10 LONGITUDINAL SECTION OF DISTRIBUTOR MECHANISM DRIVE MOTOR AND GEAR IN TURBINE SHAFT (INDIVIDUAL ELEMENTS)

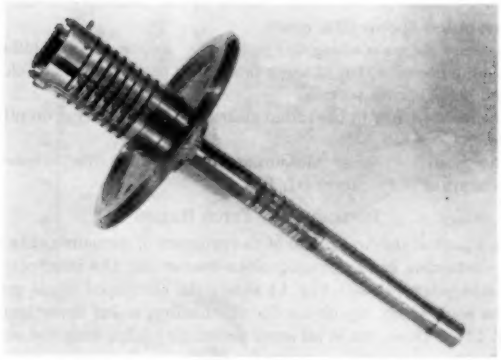


FIG. 10(a) DETAILS OF SHAFT CONTAINING MEASURING DEVICES

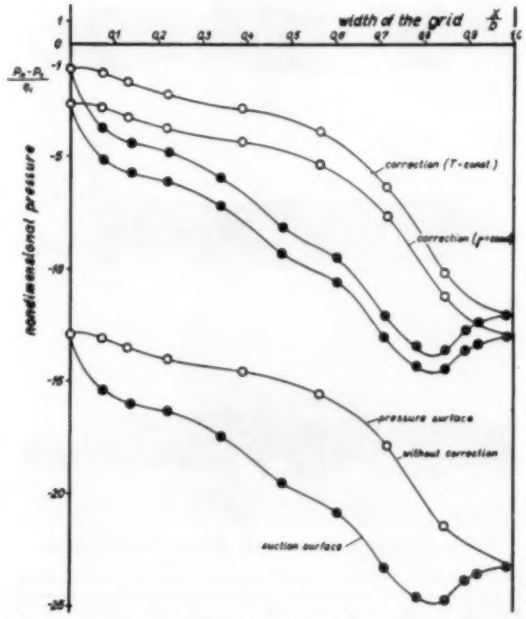
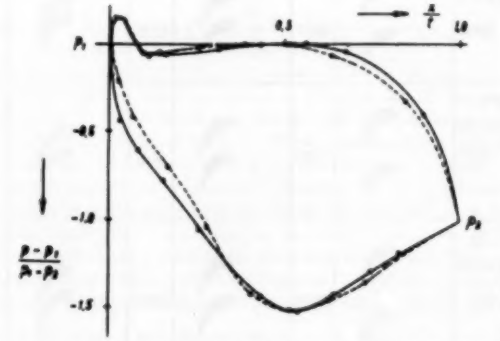


FIG. 11 INFLUENCE OF CORRECTION FOR CENTRIFUGAL EFFECT



——— Pressure distribution on plane grid (fixed)
——— Pressure distribution on rotating grid with guide vanes removed at about 200 rpm

FIG. 12 COMPARATIVE MEASUREMENTS WITH THE PLANE GRID

The influence of centrifugal forces in the boundary layers by means of low speeds (200 rpm).

Pressure increase along the radius, by means of axial inflow. The influence of wakes of the guide vanes by means of the thin-walled straightener section.

The leakage flow in the radial clearance, by measuring on pitch circle.

The result of these measurements indeed shows extensive agreement of both curves (cf. Fig. 12).

Influence of Pitch Ratios

As a part of the first phase of the program of measurements we are conducting basic investigations concerning the influence of different pitch ratios. Fig. 14 shows the developed blade grids of the root, pitch, tip circles for the blading so far investigated (Fig. 13). These are in all cases prismatic blades with the same blade angle β_s in all sections. The stator wheel was the same for all rotor rows. It consists of simple straight sheet-metal blades.

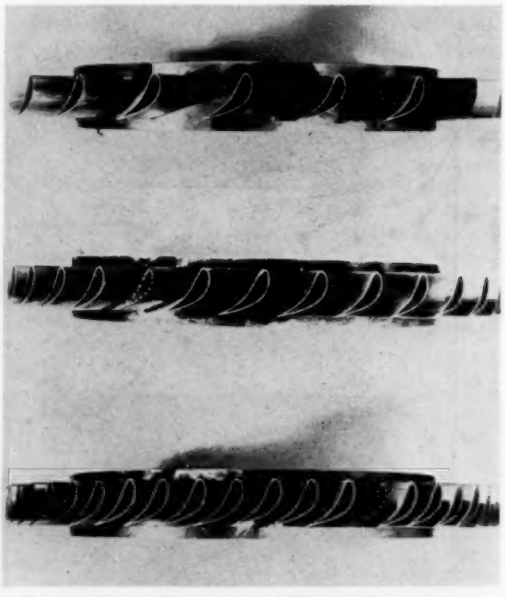


Fig. 13 Three Experimental Wheels Investigated

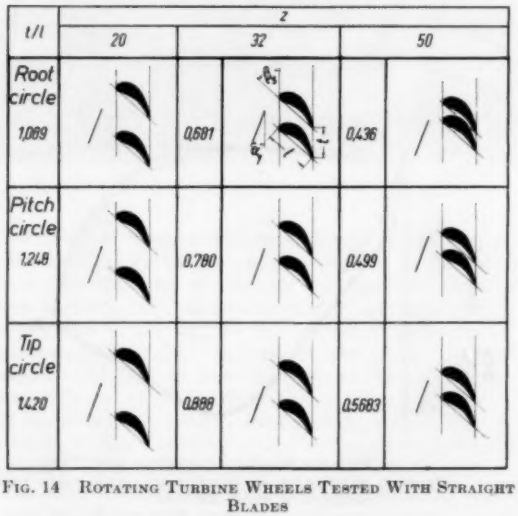
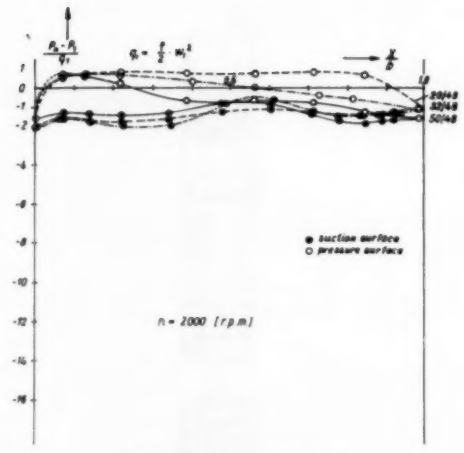
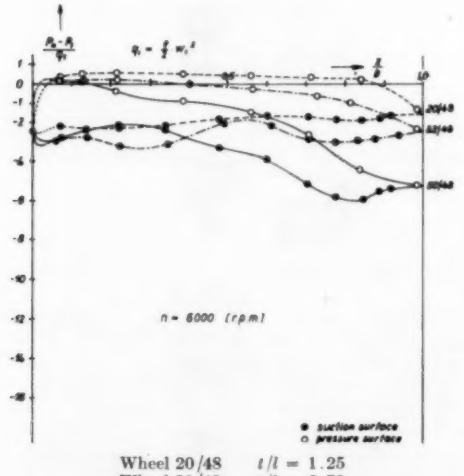


Fig. 14 Rotating Turbine Wheels Tested With Straight Blades



Wheel 20/48 $t/l = 1.25$
Wheel 32/48 $t/l = 0.78$
Wheel 50/48 $t/l = 0.50$

Fig. 15 Pressure Distributions, Section 4, n = 2000 Rpm



Wheel 20/48 $t/l = 1.25$
Wheel 32/48 $t/l = 0.78$
Wheel 50/48 $t/l = 0.50$

Fig. 16 Pressure Distributions, Section 4, n = 6000 Rpm

Figs. 15 to 17 are presented as an example of results obtained. They show the pressure distributions for the same surface of measurement of the three different wheels, plotted against the nondimensional width b of the grid. These measured results are shown for speeds from 2000 to 10,000 rpm.

The ordinates are the difference of the static pressure at any given point of the profile P_s and the static pressure ahead of the grid P_1 divided by the stagnation pressure of relative inflow velocity

$$q_1 = \frac{\rho}{2} w_1^2$$

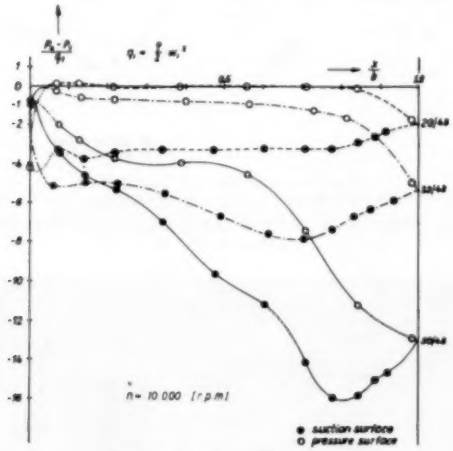
This pressure ratio should have its maximum value unity at the stagnation point. Correspondingly, at the trailing edge of the blade there should appear the entire local pressure drop for the

section of measurements. The pressure measured by the bore at the trailing edge of the blade, however, is a little lower than the static pressure P_2 measured by a static-pressure probe slightly behind the grid.

It is interesting that from a comparison of the stagnation pressure distributions is to be seen that the full stagnation pressure

$$\frac{P_n - P_1}{q_1} = 1$$

is not reached at higher speeds. This may be caused by radial components of velocity resulting, for example, from centrifugal



Wheel 20/48 $t/l = 1.25$
Wheel 32/48 $t/l = 0.78$
Wheel 50/48 $t/l = 0.50$

FIG. 17 PRESSURE DISTRIBUTIONS, SECTION 4, n = 10,000 RPM

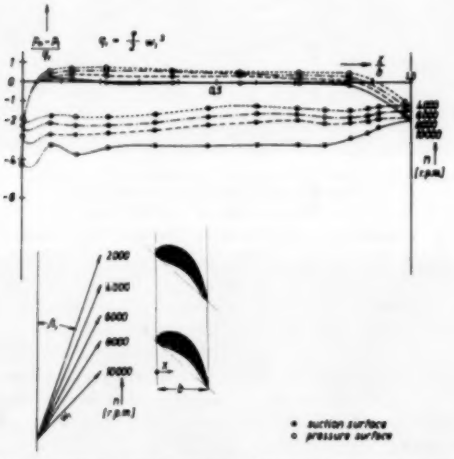


FIG. 18 PRESSURE DISTRIBUTION, SECTION 4 FOR ROTOR 20/48, $t/l = 1.25$

forces in the boundary layer or else by inaccuracies of the measurements of the static-pressure probes ahead and downstream of the rotor wheel. It was found that the shape of the curves near the leading edge could not be determined accurately since there had not been enough room to provide for a sufficient number of points of measurements around the nose. Deviations at the trailing edge may be explained by flow around the trailing edge.

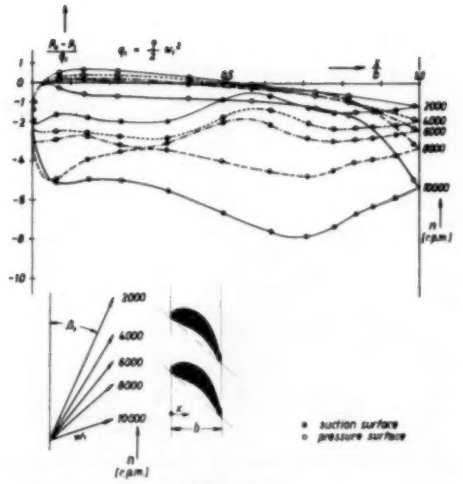


FIG. 19 PRESSURE DISTRIBUTION, SECTION 4 FOR ROTOR 32/48, $t/l = 0.78$

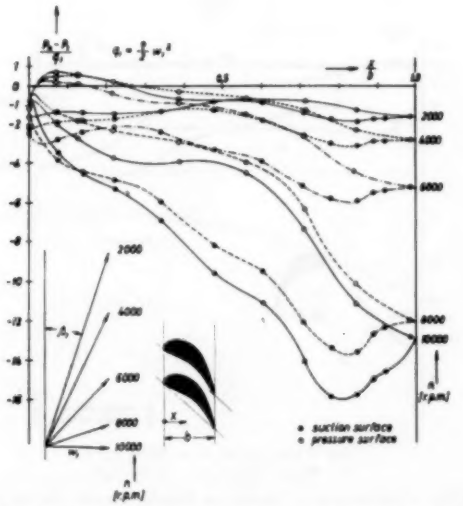


FIG. 20 PRESSURE DISTRIBUTION, SECTION 4 FOR ROTOR 50/48, $t/l = 0.50$

Figs. 18 to 20 show the same curves plotted versus speed as parameter, each for constant number of blades.

For comparison of the pressure distributions at the same relative inlet angle β_1, there are shown in Figs. 21, 22, 23 the pressures taken at radial Section 4, for two wheels identical except for a different number of blades. These measurements show the effect of closer blade spacing upon the conversion of pressure.

Figs. 24 and 25 show isometric presentations of the pressure distributions of the grids with 20 and 50 blades in all seven sections of measurement located along the blade height.

The width of the profile b was chosen as a reference line. In this case also the static-pressure difference $P_n - P_1$ was computed and rendered nondimensional with respect to the velocity pressure of the relative velocity at inlet $q_1 = \frac{\rho}{2} \cdot w_1^2$.

The tests just described were carried out with a turbine unit of the following dimensions:

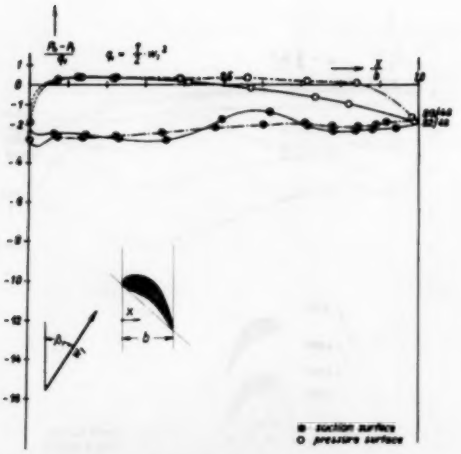
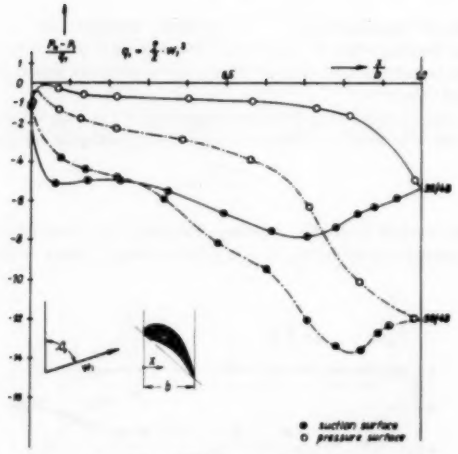
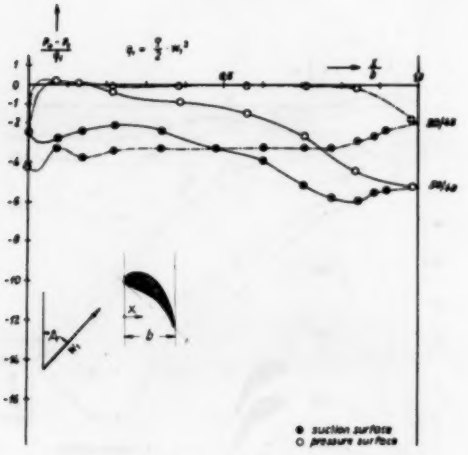


FIG. 21 COMPARISON OF PRESSURE DISTRIBUTIONS, SECTION 4 FOR EQUAL ANGLE β_1



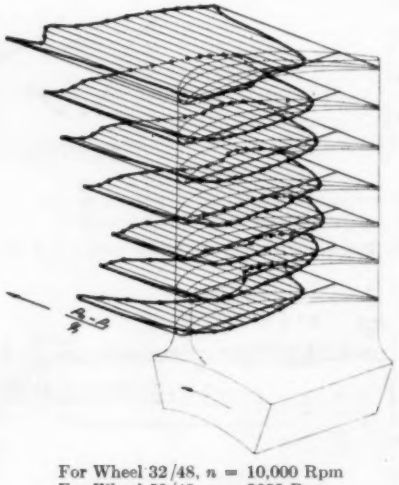
For Wheel 20/48, n = 10,000 Rpm
For Wheel 50/48, n = 6000 Rpm

FIG. 23 COMPARISON OF PRESSURE DISTRIBUTIONS, SECTION 4 FOR EQUAL ANGLE β_1



For Wheel 20/48, n = 8000 Rpm
For Wheel 32/48, n = 4000 Rpm

FIG. 22 COMPARISON OF PRESSURE DISTRIBUTIONS, SECTION 4 FOR EQUAL ANGLE β_1



For Wheel 32/48, n = 10,000 Rpm
For Wheel 50/48, n = 8000 Rpm

FIG. 24 CORRELATION OF PRESSURE DISTRIBUTION ALONG LENGTH OF BLADE, FOR ROTOR WHEEL 20/48 (t/l PITCH CIRCLE = 1.25)

Height of blades.....h = 41 mm
Pitch-circle diam.....D_m = 283 mm

hence $\frac{h}{D_m} = 1{:}7$; $\frac{r_i}{r_a} = 0.75$

Depth of profile (chord)...l = 35.6 mm
Blade angle.....β_s = 48 deg
Blade-inlet angle.....β_1 = 84.0 deg
Blade-outlet angle.....β_2 = 20.5 deg

The profile is a reaction profile. Measurements were obtained for speeds up to 12,000 rpm, which corresponds to a blade-tip velocity of 203 m per sec. The pressure ratios were chosen such that the Reynolds number was kept constant at

$$1.5 \cdot 10^5 < \mathrm{Re} = \frac{w_2 \cdot l}{\nu} < 2 \cdot 10^5$$

where w_2 = relative velocity at outlet. The Mach number was less than 0.5.

The results presented are examples illustrating the relative variations of the pressure distribution with blade pitch.

Since in the establishment of these curves measured data obtained with fixed probes upstream and downstream of the rotating grid were used, they are of the same accuracy as any information obtained from the readings of probes, including those for fixed-blade rows.

Experimental work is at present under way to eliminate as much as possible the influence of the probes for the particular case of our setup. The inaccuracies caused by them are not an inherent feature of the method for the determination of the pressure distribution on rotating grids presented here. A precise evaluation of all details of the data on hand should be deferred to the time when these special effects have been clarified and when more measurements are on hand for correlation.

A closer interpretation of all details of the measurements

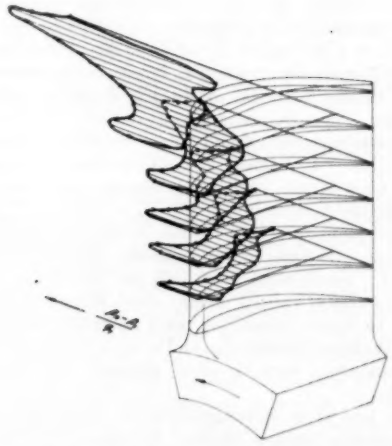


FIG. 25 CORRELATION OF PRESSURE DISTRIBUTION ALONG LENGTH OF BLADE, FOR ROTOR WHEEL 50/48 (t/l PITCH CIRCLE = 0.5)

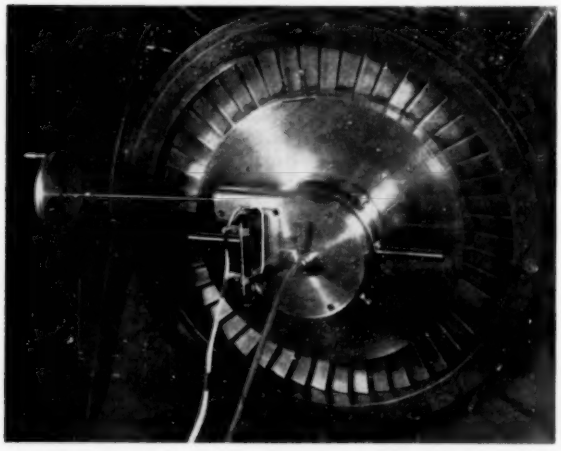


FIG. 27 OVER-ALL VIEW OF THE DEVICE FOR THE ADJUSTABLE ROTATING PROBE

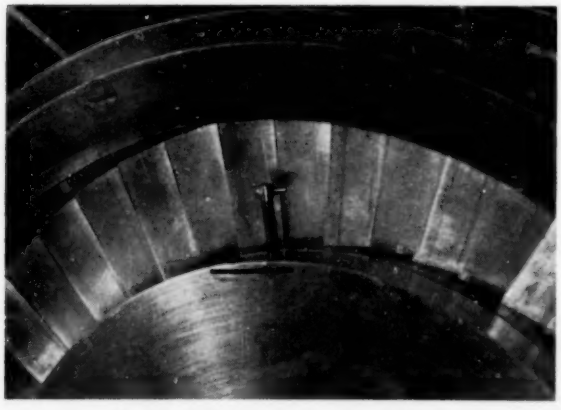


FIG. 28 TIP OF THE ADJUSTABLE PROBE

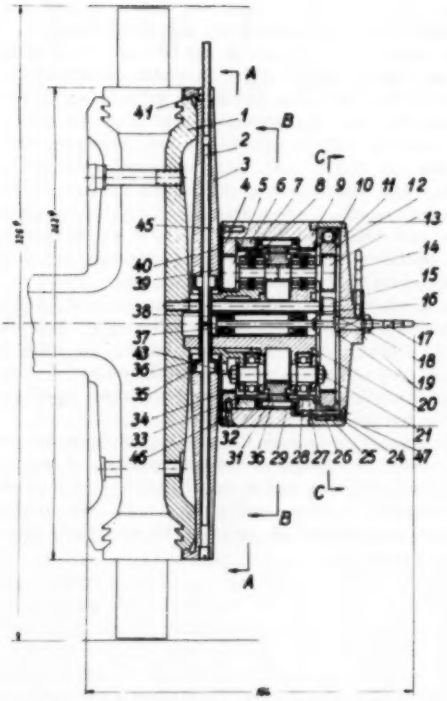


FIG. 26 ADJUSTABLE PROBE MOUNTED ON ROTOR

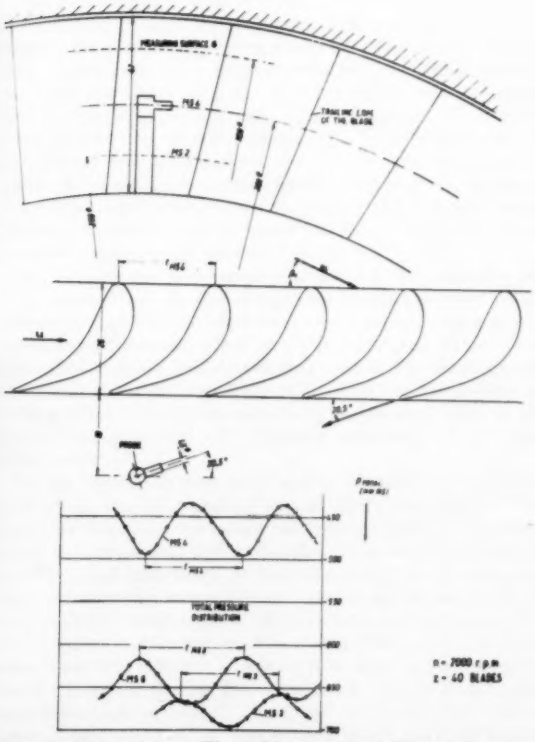


FIG. 29 RESULTS OF A WAKE MEASUREMENT WITH A ROTATING TOTAL PRESSURE PROBE

taken so far will not be attempted until a greater number of measured results are available for mutual correlation.

The experimental equipment will be augmented further by a rotating wake-traversing probe adjustable circumferentially over approximately two pitch distances. It is shown in the design drawing, Fig. 26. An attempt will be made to measure with this probe the wake troughs of the total-pressure curve and to determine in this manner local losses and their contributing causes. The picture shows, on the left immediately adjacent to the turbine wheel, a disk from which the probe projects.

The disk rotates with the rotor, but it can be moved peripherally relative to the rotor during operation by means of a stationary worm and a rotating planetary gear.

Fig. 27 shows the arrangement of the probe of 1.5-mm inside diameter. Its peripheral position is indicated by a potentiometer. Fig. 28 likewise shows the probe behind the rotor blading, oriented in its threaded mounting to the estimated direction of flow at discharge (20.5 deg). To alter the radial position of the tube the turbine needs to be stopped.

The measurements with the rotating total-pressure probe, recently begun, are aimed at the determination of total pressure losses along the span and across the pitch. An example of results obtained is shown in Fig. 29, namely, the total pressure distribution above outside pressure, 16 mm axially downstream of the rotor trailing edges, on 3 different radii, plotted against developed circumferential distance.

The pitch of wakes equals the blade pitch. The wakes of sections 6 and 2 are displaced peripherally with respect to each other because of angular deviation and because of a slight slant of the trailing edge. The centers of gravity are on a radius. Total pressure losses in sections 6 and 2 are considerably greater than on pitch circle. A detailed account of this investigation of losses will have to wait until measurements covering the entire radial height of blades are on hand.

It appears unnecessary to discuss further the purposes and potentialities of the experimental apparatus described as these are self-evident. Worth-while objectives are systematic measurements on profiles of various designs which differ, e.g., primarily in respect to shape parameters, such as median line, camber, thickness, thickness distribution, and so on, and further, in respect to grid parameters, such as, pitch, fanning, blade orientation, and the like.

This can lead to an investigation of the properties of various blades, including twisted blades. In conjunction with total-pressure measurements ahead of and downstream of the grid there exists further the possibility of facilitating the establishment of an analysis and itemized account of losses for blades of rotors with accelerating and retarding grids.

Operating Experience and Design Features of Closed-Cycle Gas-Turbine Power Plants

By CURT KELLER,[1] ZURICH, SWITZERLAND

This paper is the author's third progress report in the United States on the AK closed-cycle gas turbine. The first comprehensive presentation in 1945 (1)[2] dealt with the basic ideas and theory, the first 2000-kw experimental plant in Zurich, and future prospects, including the use of other gases than air, such as helium, nitrogen, or carbon dioxide. The second paper in 1945 (2) gave more details and design features of different components, with special reference to machines and air heater of the first industrial 12,000-kw oil-fired plant for St. Denis and Dundee which were in the final phase of erection at that time. Projects for coal-fired plants also were discussed. This third presentation summarizes the achievements during the past five years. They are based on operation and new design experience and show the marked improvements obtained by simplifying the closed-cycle system components while keeping its basic properties. This development led the way to economical solutions in the different fields of application. Work was concentrated mainly on power-station sets up to 15 mw and for different fuels as well as on ship-propulsion plants (3). The recent accelerated development of atomic power brings a new incentive to the closed-cycle gas turbine using gases other than air as the working medium and reactor coolant.

General Experience With Pilot Plants

BY THE end of 1955, 14 closed-cycle gas-turbine power plants ranging from 700 to 12,000 kw were in operation or being built by Escher Wyss and its different licensees (4). Among them are one 6600 and two 10/12,000-kw new coal-fired sets. The outstanding event of the past year was the successful start of the world's first industrial pulverized-coal-fired plant.

Generally speaking, we can state that during the entire development period no basic change of system layout proved to be necessary. Working pressures and temperatures as well as regulating means are still the same as foreseen in our first studies 15 years ago. However, practical realization of the first pioneer plants brought up a number of engineering problems and difficulties which, while having nothing to do whatsoever with the principle itself, enforced long delays on us. Strangely enough, practically no difficulties arose in any high-temperature-region components of the plants such as air heater and turbines. Not a single tube in any air heater has been damaged in normal service nor any high-temperature blading or shaft. Our difficulties occurred from so-called normal components such as auxiliaries, bearings, shaft vibrations, piping-expansion problems, compressor blades, electrical equipment, and so on. Such matters are treated in the following as they may show the difficulties which had to be overcome in bringing the first plants into reliable industrial service.

The first two plants, of 2000 kw at the Escher Wyss works and 12,000 kw in St. Denis for the EDF, were both designed, manufactured, and mounted entirely by Escher Wyss during war and after-war times. Impossibility of free choice of material and auxiliary equipment at that time was naturally a great drawback. Knowledge for many essential components, for instance on heat resisting steels, high efficiency axial compressor blading, and extended heat-transfer surface, was not yet very advanced.

We managed somehow to get through the war with all the tests and the pilot plant proved to fulfil all the expectations. For long periods, when electricity restrictions existed in Switzerland, the plant was the only energy source for the works. An over-all efficiency of 32 per cent at full load and the very flat efficiency curve over a big load range were the most striking things for the engineer. A total of 6000 industrial service hours with this test plant were run without significant trouble. We want to point out especially that the air heater, which was of quite advanced design, stood all services without any trouble. Not one single tube failed and this is true for all air heaters built up to now. This component of the closed-cycle plant which was often regarded as most questionable, turned out to be most reliable.

The regulating procedure of closed-cycle plants is quite simple by combining pressure-level variation with by-passing the compressor. No valve or regulating device is located at the machine itself. Inlet, outlet, and bypass valves for the working medium are in the cold region (Fig. 1). This is an important advantage from the operating point of view. The set answers very quickly to load alterations. This fact was apparent in the operation of the first 2000-kw test plant and was described in detail in an earlier paper. Even when disconnected from the general grid, the speed of the small 2000-kw set did not vary more than 2 per cent when load was lowered quickly by about 500 kw. A quick shutdown produced a speed rise of less than 4 per cent.

Fig. 2 shows the corresponding behavior of the 12,000-kw St. Denis plant. Even at very abrupt load changes, the working air temperature at turbine inlets varies very slowly, owing to the accumulating favorable effect of air-heater tubes. Many shutdown tests from full load showed that even in this complex plant speed increase is less than 5 per cent. This low value is due to braking effect of the compressor being on the same shaft as the turbine and generator. The modern one-shaft sets are even easier to control.

For St. Denis a top pressure of 50 atm (700 psia) and a pressure ratio of 10 with intermediate heating cycle was chosen. The reason for this choice was mainly the fear that with lower pressure ratios the recuperator and cooler dimensions would become too great with high mass flow. At that time no appropriate extended surface was available and heat-exchanging apparatus had to be designed with straight tubes. Furthermore, at that period, air-heater design problems were overestimated because of unknown nonsymmetrical radiation effects on tube walls at high temperatures. Therefore the St. Denis air heater was built as a pure convection-type apparatus. In order to reduce the surface, supercharged combustion was chosen which entailed greater complication because of the necessary charging set.

---

[1] Director of Research and Development, Escher Wyss Ltd. Mem. ASME.

[2] Numbers in parentheses refer to the Bibliography at the end of the paper.

Presented at the Gas Turbine Power Division Conference, Washington, D. C., April 16–18, 1956, of The American Society of Mechanical Engineers.

Note: Statements and opinions advanced in papers are to be understood as individual expressions of their authors and not those of the Society. Manuscript received at ASME Headquarters, March 29, 1956. Paper No. 56—GTP-15.

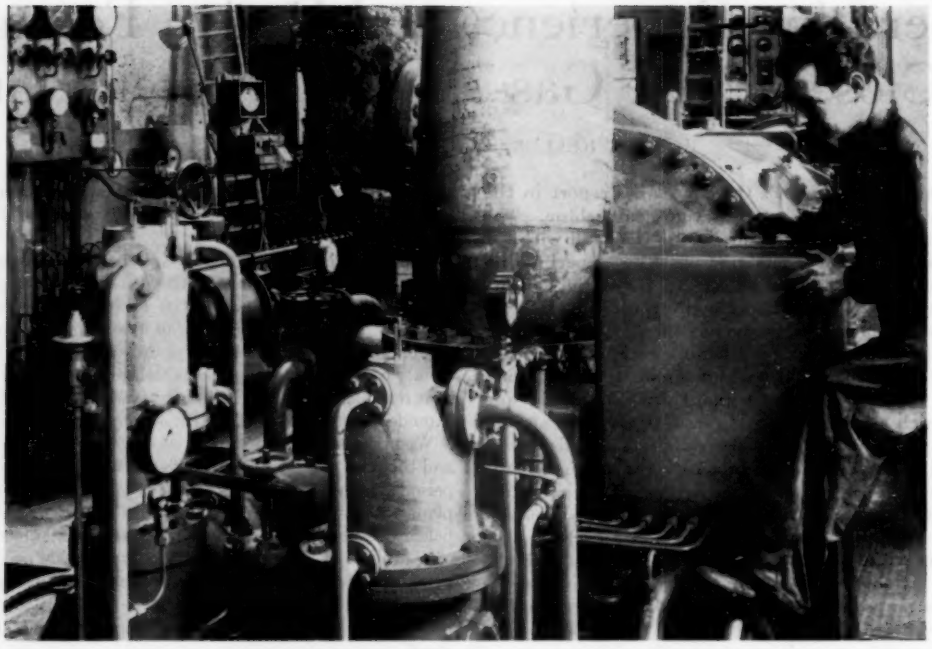


FIG. 1 COMBINED BYPASS AND INLET VALVE FOR CLOSED-CYCLE PLANT
(Whole regulating device is in covered box near engineer.)

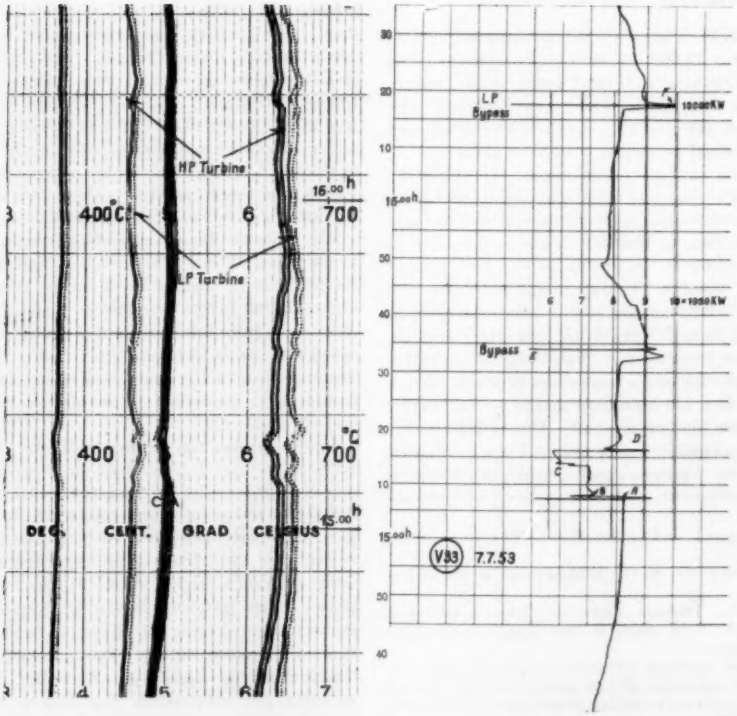


FIG. 2 SECTION OF LOAD AND TEMPERATURE DIAGRAM OF 12,000-KW PLANT DURING REGULATING TESTS

In the final uninterrupted continuous run of 700 hr in 1955, most valuable measurements and tests were made with runs up to full load and top temperature of 675 C (1250 F). An inspection of the machine and air-heater system after more than 3000 hr of running time showed clean machines, recuperators, and piping. Also the outside surfaces of the air-heater tubes were in very good condition and no corrosion was experienced.

In the following we will mention some of the annoying difficulties which occurred at St. Denis. None of them related in any way to the air-heater performance or with the high-temperature parts. The plant was shut down recently after having proved to be quite satisfactory in mechanical behavior. At present some machines are undergoing general alterations. Based on new experience, some of the old blading for compressors and turbines, which were damaged by earlier accidents, will be replaced by improved blades.

In order to illustrate our troubles a list of some events which occurred at St. Denis since starting up in 1952 follows:

Primary incorrect insulation of a double-wall high-temperature inlet pipe to turbine caused some turbine vibrations because casings were displaced by unequal temperature elongations.

After a 200-hr run in July, 1952, a slight vibration of the high-pressure (H-P) radial compressor was noted. Shortly afterwards the H-P turbine set speeded up and consequently the quick-closing device of this group acted automatically and the whole plant was shut down as foreseen. An examination of the H-P compressor rotor showed that the blades of one runner wheel had been broken away. The detailed inspection proved that some stationary diffuser blades from the preceding stage had been welded unsatisfactorily. Because of the stream forces acting on these deflector blades during normal operation the welds became loosened and pieces of the blades were thrown into the following runner wheel, causing the damage.

The charging and leakage compressors of the unlubricated piston type were unreliable at first and had to be rebuilt by the manufacturer. Now we use rotary Lysholm-type compressors.

At the start the electrical equipment of the plant caused some short circuits in the auxiliary motors, and various faults in the electrical relays hindered proper functioning of the safety devices. Hence, at a shutdown test, the electrically driven auxiliary oil pump did not function and some bearings of the machines were damaged.

The main cooling-water electric motor once caught fire from some unknown cause during normal service and required a long time for repair.

The cooling-water conditions at St. Denis were unfavorable during summer months as temperatures went up unusually high (to 90 F) and the water was very dirty. The coolers required frequent cleaning.

At first, high-pressure and low-pressure sets were running quite satisfactorily, but after various necessary dismantlings of the machines, the running behavior got gradually rougher after realigning the shafts. It took a long time to find out that both groups were running too close to the critical speed, because the theoretical calculation did not indicate this as the cause. Model tests showed that assumptions of the calculation were incorrect and that the critical speed was too close to the running speed. When the shaft and bearings were aligned not too correctly, vibrations were avoided by additional bending but the more accurate the alignment, the worse the running became. Fortunately, by removing only one bearing in each L-P and H-P set, the critical speed of the shaft could be entirely corrected and since then the running behavior is perfect.

Measurements show that the character of the flat efficiency curve corresponds to the expectations (Fig. 3) and that the output is proportional to the pressure level.

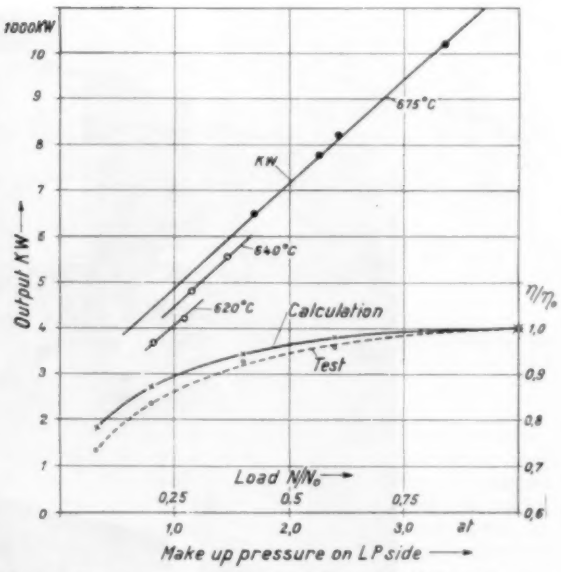


FIG. 3 EFFICIENCY AND OUTPUT OF FIRST 12,000-KW PLANT AT DIFFERENT PRESSURE LEVELS (LOADS)

The first closed cycle-gas turbine power plant in Scotland, at Dundee, built by John Brown & Co. Ltd., differs not too greatly from the St. Denis plant in its components. It is also a complex two-shaft group. The air heater with supercharged combustion chamber is built for oil firing but is much smaller than the corresponding St. Denis apparatus. A cross-flow combustion-type air heater is used. This plant was started in the Fall of 1954, without too great difficulty. However, after less than 100 hr of service the alternator caught fire and was heavily damaged. The reasons for this accident are not clear. This failure delayed operation of the plant for a year. Later a blade damage in one of the axial compressors occurred, similar to a corresponding failure in the St. Denis plant. As the forces in compressors running at high density are greater than in normal atmospheric or low-pressure compressors, improved structural design of stator and runner-blade roots is necessary. Work of altering the Dundee plant is still under way.

In addition to the 12,000-kw Dundee plant, John Brown Co. has pioneered and completed successful development work and tests with coal and peat-fired air heaters (Fig. 4). This work is sponsored by the North Scotland Hydro-Electric Board as well as by the British Coal Board. A 2000-kw peat-fired industrial plant incorporating this air heater and a "Tuco" set similar to Fig. 7 are under erection in the North of Scotland, as well as a 2000-kw set with a coal-fired air heater for Scotland.

John Brown uses a vortex slag-tap combustion chamber for its solid-fuel-fired heaters while in the Continental development normal conventional pulverized-coal burners with granulating combustion chambers are used. Furthermore, John Brown has built the first 700-kw closed-cycle plant using waste gases from a gas works. The start of operation of this plant in Coventry was delayed until the middle of 1955, owing to building restrictions in England. This plant is undergoing extensive service tests now.

It is believed that the waste-heat recovery from all sorts of chemical plants offers a promising field of application even for smaller outputs, because efficiencies of gas turbines can be kept higher than with the corresponding small steam turbines, and high temperatures can be used without the complications involved

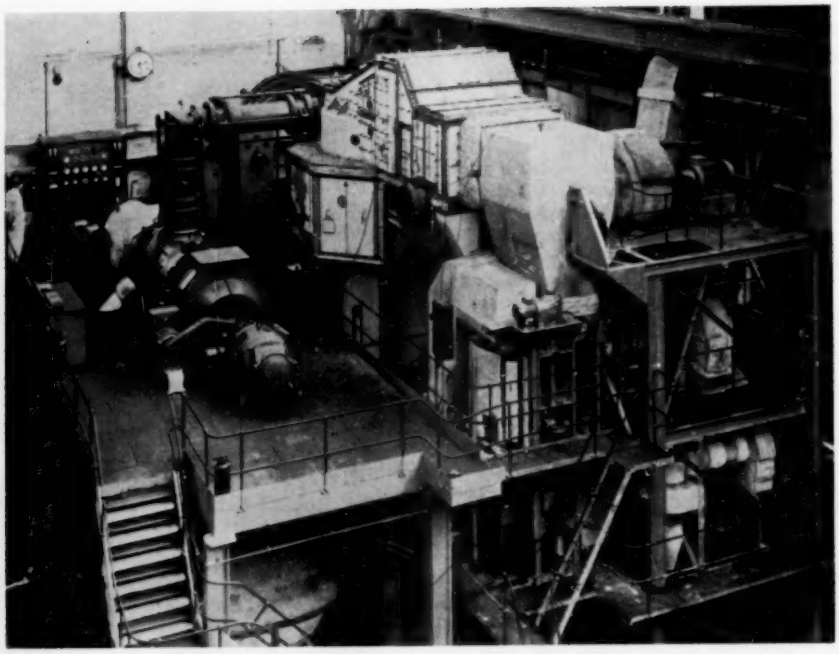


FIG. 4(a) JOHN BROWN'S TEST PLANT FOR PEAT AND COAL-FIRED AIR HEATERS (1955)

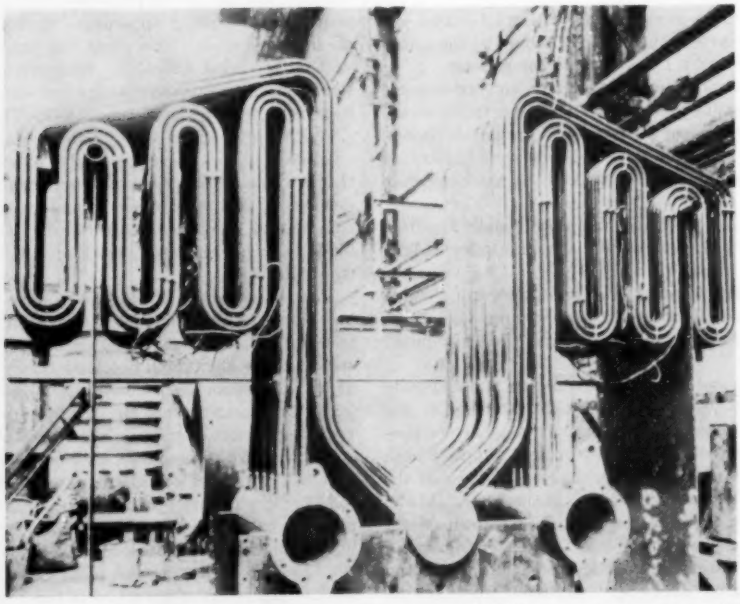
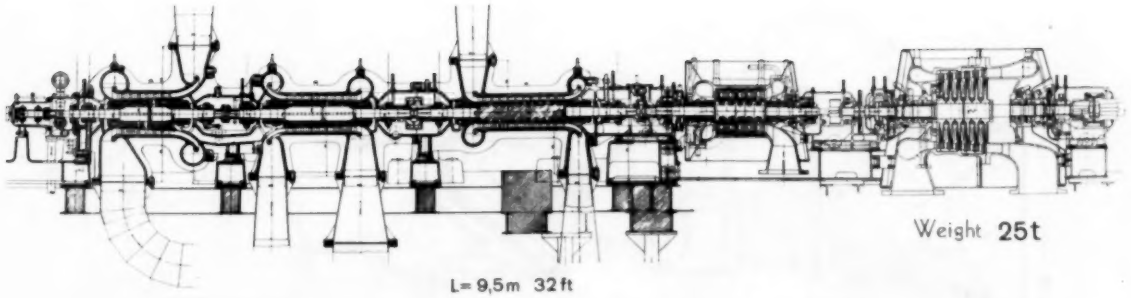


FIG. 4(b) JOHN BROWN'S TUBE SYSTEM OF EXPERIMENTAL AIR-HEATER FOR SOLID FUEL (1950)



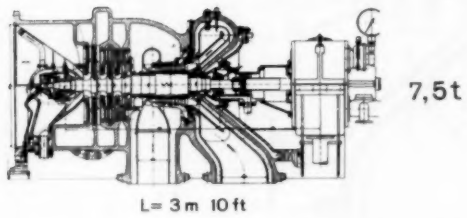


FIG. 5 COMPARISON OF 2000-KW CLOSED-CYCLE MACHINERY OF TEST PLANT (1939) WITH TODAY'S MACHINE SET OF THE SAME OUTPUT

with steam turbines. Power production per calorie of waste heat is high. If necessary, steam production for process steam can be combined with the closed-cycle plant.

ADVANCED MACHINERY DESIGN

As time went on, it became more and more obvious that the earlier complex plant design with different compressors and turbines or 2-shaft arrangement was too complicated as well as too costly and uneconomical. It was clear also that the first big power plant of 10/12,000 kw, designed and built for the EDF (Electricité de France) in Paris around 1946, was developed along too academic lines.

The progress in simplified design between the first 2000-kw test plant in 1942 and today's design is best demonstrated by Fig. 5, giving both machine sets to the same scale. The old turbine which ran so satisfactorily is the first gas turbine working with 700 C (1300 F) in industrial service. It now occupies a place of honor at the permanent exhibit of pioneer machines at the Deutsche Museum in Munich.

The new Tuco machine proved to be very successful from the start. Only 4 hr after the first run, in 1955, we could connect the set to the grid and pick-up load. After 500 hr of thorough tests this first machine was shipped to Japan. It is part of the first closed-cycle 2000-kw oil-fired plant in that country by Fuji-Denki which started operation in December, 1955. Fig. 6 shows this plant on site.

We have concentrated our later design and development studies on three standard sets (2000–3000, 5000–6000, and 10,000–15,000 kw). The two newly ordered 10/12,000-kw sets are being built as Tuco sets, where compressor and turbine are combined on one rotor, similar to Fig. 7(*a*). Leakage losses are reduced because for this arrangement of the compressor and turbine no high-pressure gland is necessary. All these machines can be used for oil, gas, or coal-fired air heaters.

In the near future development of additional standard sets above 12,000 kw is foreseen. As mass flow of working fluid is becoming larger, the best solution for the machine will then be axial-flow type machinery. Fig. 8 shows a new compact machine design of 15 mw suitable also for ship propulsion. Here again no high-pressure gland has to be provided, only the low-pressure sides of the machines have to be tight against the atmospheric surroundings. The design pressure for this set is 35 atm. There is no reason why the pressure level should not be increased. As the diameters of the machines are small, the pressure of 70 atm would result in a useful output of the machine as per Fig. 8 of approximately 30 mw. The free choice of the pressure level is one of the most important design advantages offered by the closed-cycle plant in order to cut dimensions.

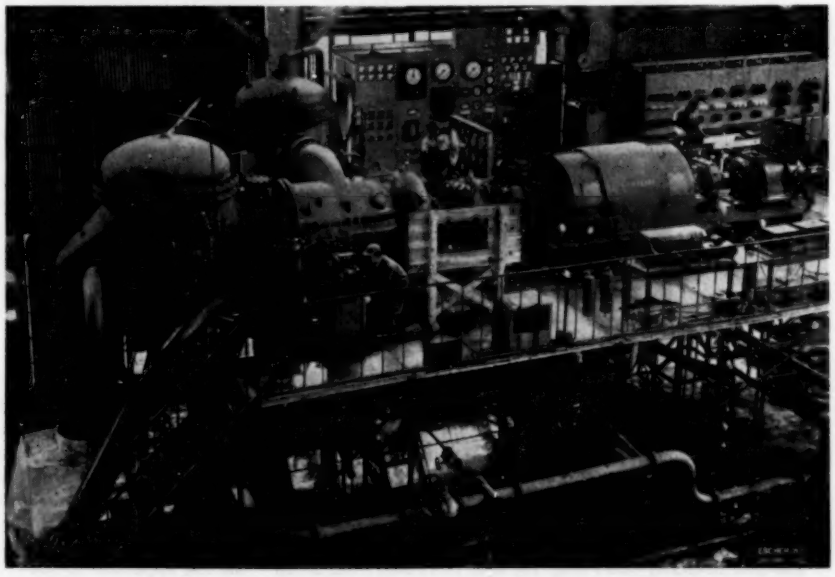


FIG. 6 FIRST JAPANESE CLOSED-CYCLE GAS TURBINE ON TEST BED

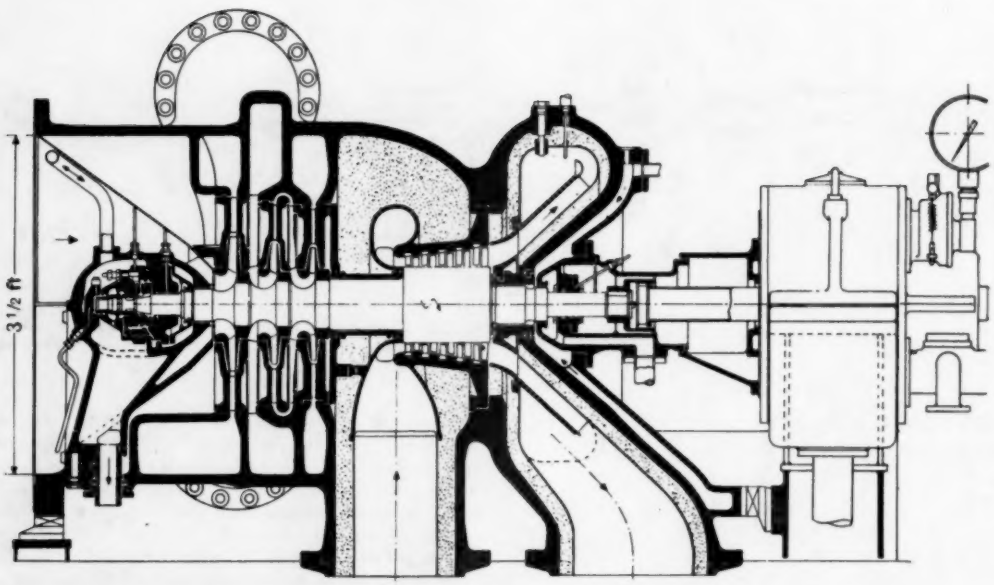


FIG. 7 CROSS SECTION OF NEW TUCO MACHINERY SET. TURBINE AND COMPRESSOR ON SAME SHAFT

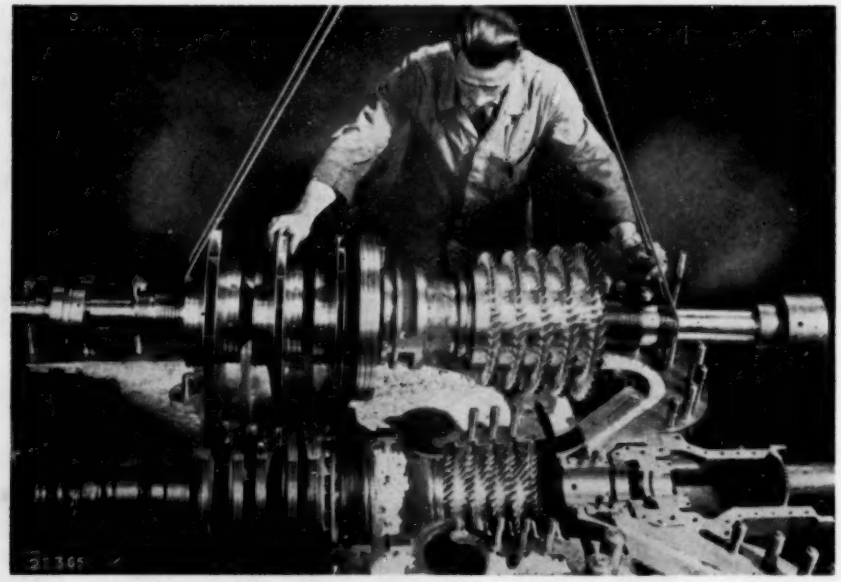


FIG. 7(*a*) ROTOR OF 2000-KW TUCO MACHINE

No obstacle is seen in principle, therefore, from the design point of view in building closed-cycle gas turbines for 50 to 100 mw. The rotor dimensions of a 50-mw set with 35 atm working pressure (which can deliver 100,000 kw when the design pressure is boosted to 70 atm) are shown in Fig. 9.

AIR-HEATER PROBLEMS

The first air-heater tubes (Fig. 10) were fabricated from 25/20 chrome-nickel heat-resisting (Thermax 11) steel strips which were bent and welded longitudinally. At that time (1937) no reliable figures on long-time creep tests were available so we conducted an investigation in our metallurgical laboratories. The most interesting creep tests were at 720 C under stresses and temperatures which occur in air-heater tubes (1400 F, 500 psia, Fig. 11). The results of a temporary involuntary overheating are shown f.i. in these curves. An unforeseen temperature rise of 100 C in the test apparatus to 820 C for 4 hr, when the tests already had been under way for 15 days, did no harm to the test specimens. It

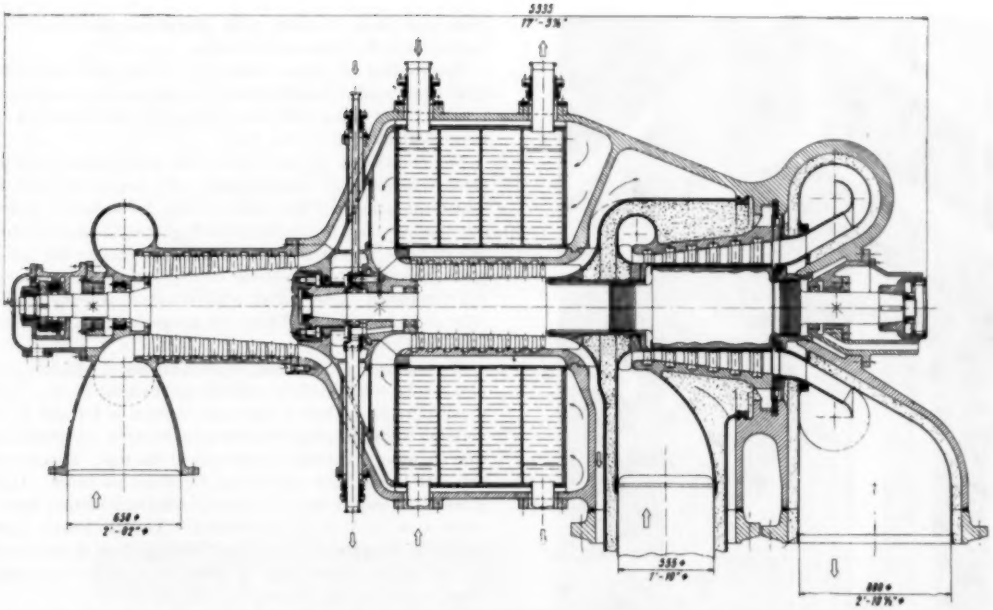


FIG. 8 CROSS SECTION OF ADVANCED DESIGN 15–20-MW CLOSED-CYCLE MACHINE GROUP. BUILT-IN INTERCOOLER

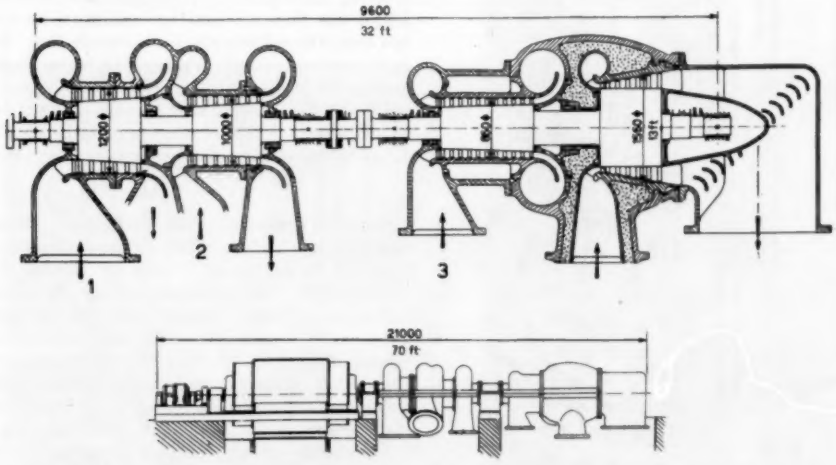


FIG. 9 STUDY FOR 50-MW CLOSED-CYCLE MACHINE GROUP DEMONSTRATES SMALL ROTOR DIMENSIONS

1 From precooler, LP compressor inlet.
2 From intercooler I, IP compressor inlet.
3 From intercooler II, HP compressor inlet.

is evident that this high overload only produces a more rapid elongation. When normal test temperature is resumed, creeping continues normally on a somewhat higher level. No crack occurs, but rupture time is slightly reduced. This test shows that safety of air-heater surfaces is quite good even at an unexpectedly high degree of overheating because tube stresses are very low compared with blade stresses in open-cycle turbines. Long-time stress-rupture tests for different materials at the Escher Wyss laboratories run up to 45,000 hours. About 450 test specimens are undergoing tests. Short-time rupture stress is about 10 times higher than normal working stress and can be withstood for many hours.

Another series of material tests is being conducted with real tube samples under internal pressure. With such tests we have now reached 30,000 hours at 750 C and 50 atm (700 psia) inside pressure, corresponding to full-load service. G 18 B tubes for instance showed a growth of only 2.4 per cent in diameter after 28,000 hours; type 316 tubes 2.8 per cent in 20,000 hours. Rupture does not occur below elongations of 15 to 20 per cent. Wall stress is approximately 2 kg/mm² (2800 psi). Tubes which become overheated do not crack suddenly. As Fig. 12 shows, a strong bulging action occurs many hours before cracking, which shows that the danger point is approaching. No explosion occurs.

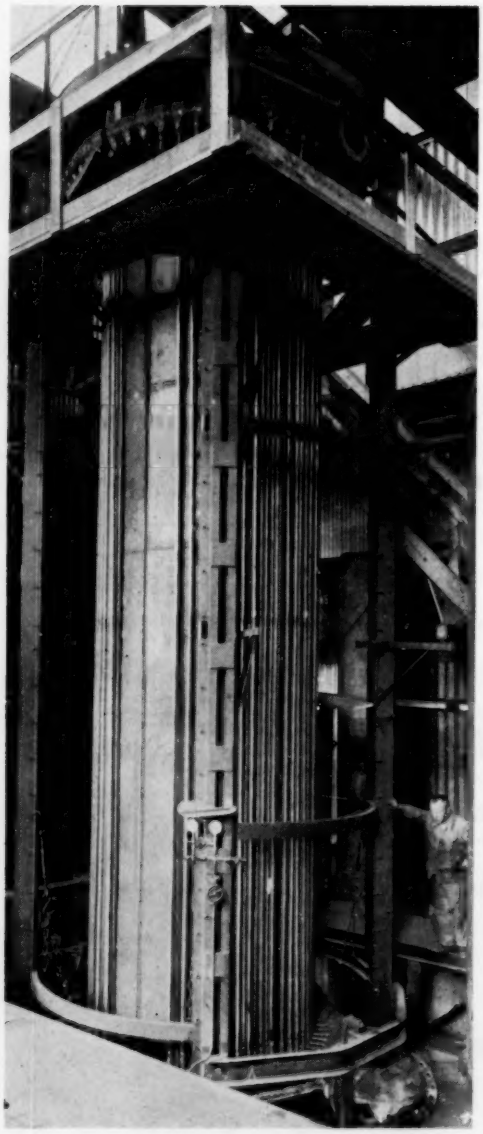


FIG. 10 FIRST ESCHER WYSS AIR HEATER OF 1939 TEST PLANT SHOWING EXTERNAL CONVECTION TUBE RODS SURROUNDING INTERMEDIATE COMBUSTION-CHAMBER WALL

[Radiation tube rows are inside intermediate wall. Combustion room diam 120 cm (4 ft), height 7 m (23 ft), specific load of combustion room 830,000 kcal/m³ (93,000 Btu/cu ft).]

Contrary to steam-boiler practice we can use small-diameter thin-wall tubes to build up the heating surface (approximately 1 in. diam, 0.1 in. wall). Drums are not necessary in modern air-heater design. Small tubes are welded in individual conical headers from which the working gas is led in a small number of collecting tubes to the final collectors (Figs. 13, 14). Each tube bundle is accessible and can be removed easily. The maximum tube-wall temperature never exceeds the working temperature of the inside gas by more than 50 C (90 F), due to the high heat-transfer coefficient inside (250 to 500 Btu/sq ft hr deg F). This has been checked with thermocouples in different air heaters as in St. Denis and Dundee.

Thin-walled air-heater tubes of ferritic and austenitic steel have been welded together and this connection has proved to be reliable, even under 1000 times repeated thermal-shock tests between 300 and 600 C (Fig. 15).

The question of chemical attack by combustion gases has been given our greatest consideration. We began our first tests on vanadium attack at the close of 1945, for instance, and pointed out the influence of additives for high-vanadium-containing fuel oils in basic patents. One recommends raising the ash-melting point, the second the delusion effect, and another evaporation of the vanadium oxide. Also sulphur attack is being studied carefully and has been found to be not dangerous for alloyed steel up to 700–750 C.

We believe that corrosion is not a chemical problem alone, but also can be combatted by suitable mechanical design. Therefore, in all our air heaters a radiation section is formed by straight vertical tubes, forming what is practically a cylindrical combustion chamber. Usually burners are at the top. This prevents ash deposits being built up on high-temperature tubes. Only when gases have cooled down do they reach cross-section tubes in the convection part of the air heater. The air heater approaches tube-still design philosophy more closely than it does that of the steam boiler, where similar conditions exist concerning tube stresses and temperatures.

The latest oil-fired air-heater design by Escher Wyss for a stationary power plant, where low space requirements are not of concern, has been built by our Japanese licensees (Figs. 16, 17). This design offers very good accessibility to the entire tube surface and should be suitable also for low-grade fuel. For marine plants much more concentrated designs with higher combustion chamber and specific tube-load figures can be accepted. For naval air heaters we have designs which require only 0.2 cu ft per hp. Such an air heater for 2 × 10,000-shp plant is being manufactured in Japan by one of our Japanese licensees (Fig. 18).

COAL-BURNING PLANTS

Since the beginning of the development Escher Wyss has concentrated much of its efforts on developing the coal-burning gas turbine. In collaboration with our licensees "Gutehoffnungshütte (GHH) and "Kohlenscheidungs-Gesellschaft" (KSG), in Germany, we have designed and built the first 2000-kw plant which went into operation in Ravensburg, Germany, in December, 1955 (Figs. 19, 20). This unit not only furnishes the entire electric power for an industrial plant of 1500 workers, but also delivers the necessary heating for the whole establishment and the workshops. Combination of power and heat production in the closed-cycle plant is very promising. Compared with the corresponding steam plant, the electric output per available heating unit is higher. Furthermore, power production and heat production are independent of each other. Warm cooling water of the intermediate compressor coolers as well as the precooler of the cycle is available without affecting the power-production cycle itself at elevated temperatures of 80–110 C (175–215 F). The over-all efficiency in such a combination is about 60 per cent; i.e., about 30 per cent of the fuel energy is transferred into electric power and an additional 30 per cent of the fuel energy is transferred in useful process heat. An approximate scheme of the Ravensburg plant is given in Fig. 21. The pulverized-coal-fired air heater stands as an outdoor unit near the powerhouse (Fig. 20) as no danger of freezing exists in the air heaters.

There is a great trend in Europe to combine power and heat production in thermal plants as a means of improving fuel economy. In many parts of Europe electric-energy costs are high (2–2½ cents/kwhr) even when produced by large steam

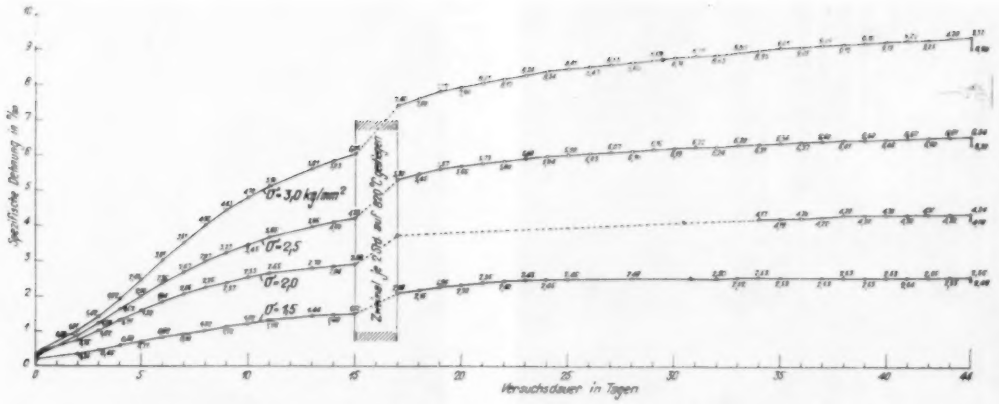


FIG. 11 CREEP TEST 1937 WITH HEAT-RESISTING MATERIAL (25/20 STEEL) (T = 720 C)

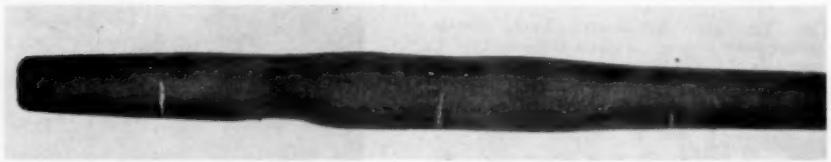


FIG. 12 CREEP TESTS WITH SAMPLES OF AIR-HEATER TUBES UNDER INTERNAL PRESSURE

FIG. 13 INDIVIDUAL TUBES OF RADIATION SECTION ARE COLLECTED IN CONICAL HEADERS

units. This offers advantages for high-efficiency gas turbines and favors decentralization of power production. The smaller self-contained power units from 5–20 mw are of interest to many countries, especially Germany where both transmission of electricity and coal-transportation costs are high. Two more coal-fired plants for isolated town works, one of 12,000 kw and one of 6600 kw, combined with remote heating for the surrounding buildings, have been ordered recently and will come into operation in 1958 as a result of the successful starting operation of the pilot plant in Ravensburg.

In Fig. 22 the layout of a 6600-kw closed-cycle plant is given. It should be noted that the first 2000-kw coal-fired air heater was based on very conservative figures in order to gather experience. Compared with the corresponding steam boilers, the specific load of the combustion chamber was very low (125×10^3 cal/m^3 = 14×10^3 Btu/cu ft). No difficulty was encountered with the coal-fired air heater. The height of the combustion chamber is dictated by the manner and timing of feeding the coal as required for perfect combustion, and depending somewhat on coal quality. Air heaters also will burn poor quality coal with as much as 35 per cent ash as well as brown coal. Burners are placed at the top of the cylindrical combustion cylinder. We realize that the first 2000-kw air heater is the lower limit of economical and technical application. Fig. 23 shows that the air heater now being built, having

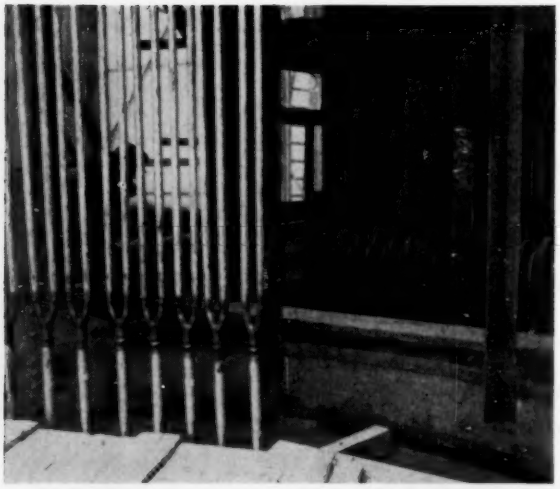


FIG. 14 DETAIL OF COAL-FIRED AIR-HEATER TUBE SYSTEM (Radiation part in foreground. Convection part in background.)

more than 3 times greater output than the Ravensburg heater, has practically the same dimensions and a simplified tube arrangement.

It should be noted that when steam is both the working and heating medium, the steam boiler must be designed for higher steam production than would be necessary for power production alone. Therefore, steam boilers for combined power and heat production get larger and more costly. The closed-cycle system provides hot water as waste heat from the cycle. Hence, the air heater for a power production plant or combined power-heat production plant, is the same. This fact helps to make such closed-cycle plants in this field economical. Modern closed-cycle plants using only moderate pressures from about 25–50 atm (350–700 psia) top pressure are much simpler in layout and operation than modern and complex steam cycles with high pressures and temperatures which are profitable only in very large units.

SHIP-PROPULSION PLANTS

As the specific fuel consumption with the closed-cycle ship plant is practically constant over a wide range of load, the closed-cycle gas turbine offers special advantages compared with other machinery for naval vessels. Special features of the ship-propulsion closed-cycle plant have been discussed in detail in a recent paper by Keller and Spillmann (7).

The ideal propeller to go with gas turbines is the automatically controlled variable-pitch propeller. Recent experiences of Escher Wyss show that such propellers can be built without difficulties up to 10,000–15,000 shp, based on the experience with Kaplan turbines. In the past two years an increasing number of such propellers was built. Prejudice seems to have been overcome by successful demonstration of a great number of recent installations, especially in France and Germany. We feel that variable-pitch propellers up to 20,000 shp can be built safely.

Another solution for the propeller drive of a gas turbine is the radial-inflow turbine. By variable guide vanes the sense of rotation of the fixed propeller can be changed. Owing to the elevated pressure level of the closed cycle, the dimensions of the power turbine can be kept small (Fig. 24). One turbine resembling the Francis-type water turbine can produce 10,000 shp at 10^4 rpm with not more than 710 mm diam. For ship plants the closed-cycle turbine is divided into a compressor-turbine group, similar to the

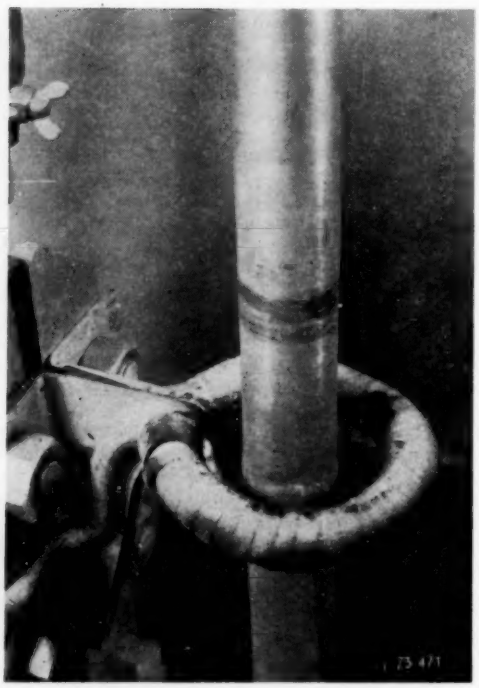


FIG. 15 WELDED AUSTENITIC/FERRITIC TUBE SECTIONS UNDER THERMAL SHOCK TESTS

Tuco design, and an additional independent propeller turbine. This second turbine receives the exhaust gas from the H-P compressor-turbine at a relative low pressure and temperature, so that no special design conditions exist.

ATOMIC POWER PLANTS

The new simplified machine designs have been developed with a view to using the same basic elements both for fossil-fuel-fired plants and future atomic-energy installations. At the beginning of our development work we pointed out the unique property of the closed-cycle system to use different gaseous working media. In 1945 a scheme for using this cycle in combination with a gas-cooled reactor was studied, but time was not ripe then for that proposal, owing to the lack of reactor technology for high-temperature reactors (Fig. 25). Several patents were applied for, using helium, nitrogen, or helium-CO_2 mixtures and proposing different working schemes. Escher Wyss is pleased to learn that recently these proposals are being reconsidered. The basic ideas were first published in the United States by R. T. Sawyer (8), but did not become widely known. An old internal Escher Wyss report written in 1945 is becoming up to date again now, and reads as follows:

The physical characteristics, for example of helium or mixtures with carbon dioxide are such, with regard to the heat transmission, that for the same losses of pressure and the same absolute pressures a heat transmission results which is about three times greater than is the case of air.

Thus, for otherwise similar conditions, the heating surfaces in a helium heater could again be considerably reduced compared to the air heater. However, if the same peripheral speeds are to be employed, light gases call for the adoption of more stages in the machines. In this connection it is worthy of note that helium, for

instance, has a sonic velocity which is approximately three times greater than that of air at the same temperatures. In principle, this would permit of the admissible circumferential speeds of turbomachines being increased, since they are dependent by no small degree on the closeness of the approach to the velocity of sound.

Moreover, the possible increase of the peripheral speeds in the event of helium being adopted, again leads to a reduction in the number of stages, so that in this connection also the constructional prerequisites are not unfavorable, which incidentally already has been shown by corresponding investigation carried out by Professor Ackeret.

If one compares the project of a gas-turbine nuclear power plant which was designed by Ackeret in November, 1945 (Fig. 26), with the modern aspects of such plants, the preview of future realization possibilities even for reactor dimensions and control

FIG. 17 JAPANESE-BUILT AIR HEATER FOR A 2000-KW PLANT, SEE FIG. 6

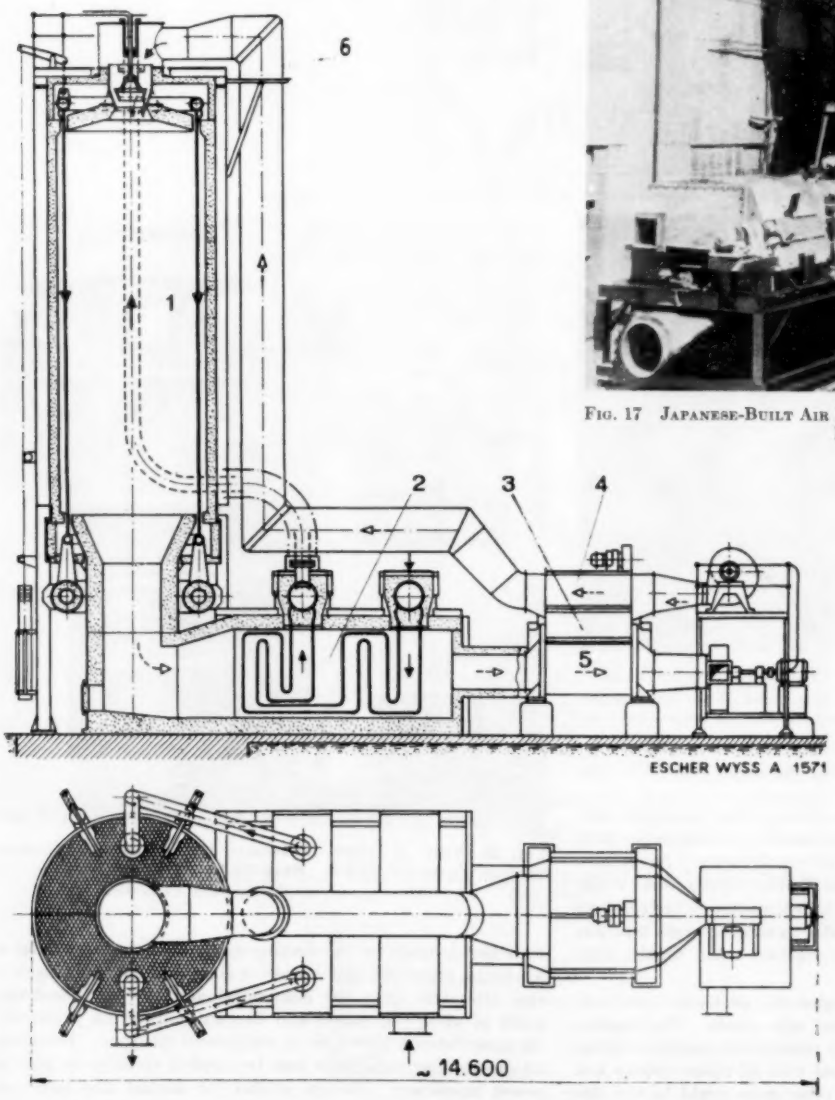


FIG. 16 CROSS SECTION OF OIL-FIRED AIR-HEATER

1 Combustion chamber and radiant section
2 Convection part
3 Combustion air preheater
4 Combustion air in
5 Combustion air out
6 Burners in top of air heater

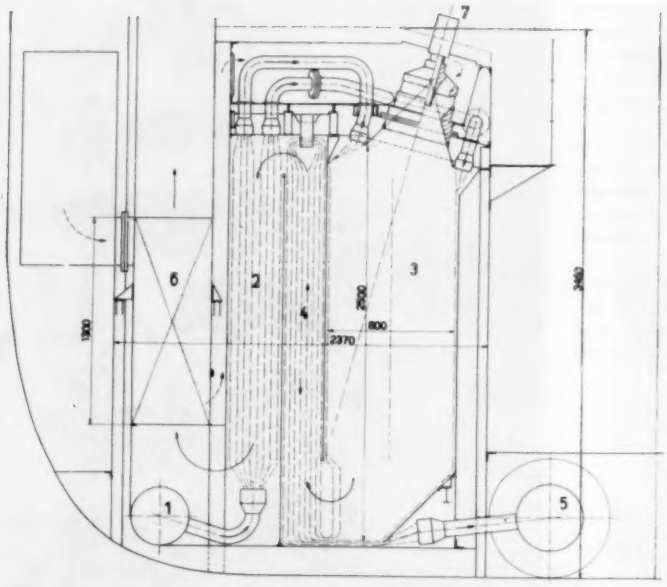


FIG. 18 CROSS SECTION THROUGH NAVAL AIR HEATER FOR 10,000-SHP PLANT

1 Air inlet from heat exchanger 400 C
2 Primary convection section
3 Combustion chamber
4 Secondary convection section
5 Air outlet to H-P turbine 675 C
6 Combustion-air preheater
7 Oil inlet for burner

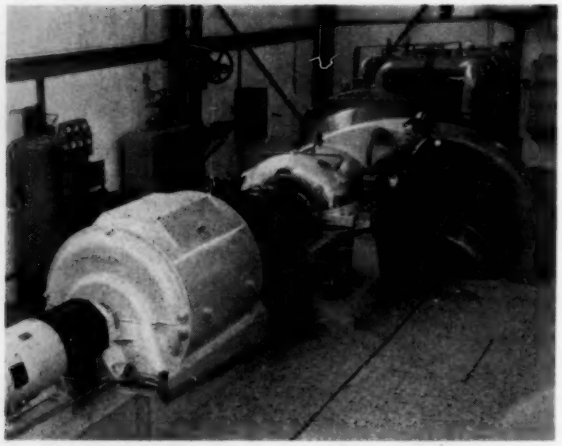


FIG. 19 MACHINE SET OF FIRST COAL-FIRED 2300-KW PLANT

[Left to right: Generator with starting motor, Tuco set (turbine-compressor set), precooler used as water heater for central heating.]

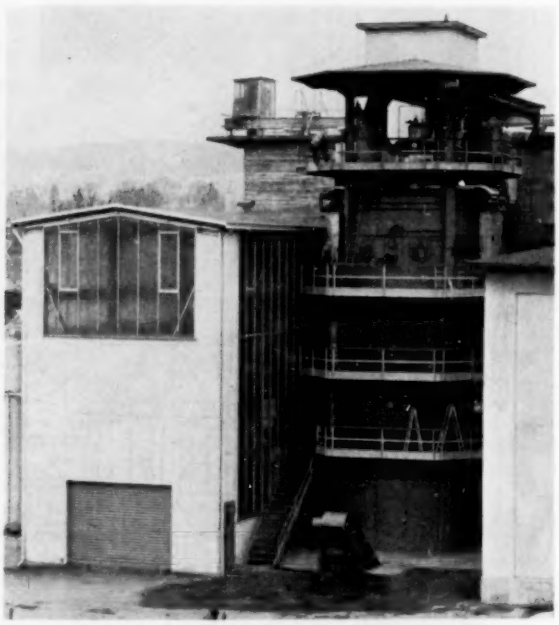


FIG. 20 VIEW OF FIRST INDUSTRIAL COAL-FIRED GAS-TURBINE PLANT—2300 KW. STARTED OPERATION IN 1956

(Left side, powerhouse; right side, air heater.)

details was not bad. We believe this document to be of historical interest to engineers.

Today's development in elevated-temperature gas-cycle reactors bears out the feasibility of such ideas by the impressive work done in the AEC-laboratories and private industry. As revealed by the recent ASME Nuclear Gas Turbine Symposium, Washington, D. C., December, 1955, (*a*) high-temperature fuel elements will be available and, (*b*) it is possible to maintain high heat flux in small-sized, low-pressure-drop reactors when highly pressurized gas is used.

We realize that a number of engineering problems have to be solved before building reliable and safe plants. The requirement of tightness for such a plant necessitates special construction, but such matters are involved with all other systems too. Helium or helium mixtures with other gases might be the ultimate best solution for the working medium, but recent studies of American physicists and engineers show that pure nitrogen is also attractive from the economical and the machine-design point of view. Machines and heat-exchangers are practically the same for compressed air or compressed nitrogen. Therefore, all air-machine experience can be applied directly to nuclear power machinery. Design studies for normal and large nu-

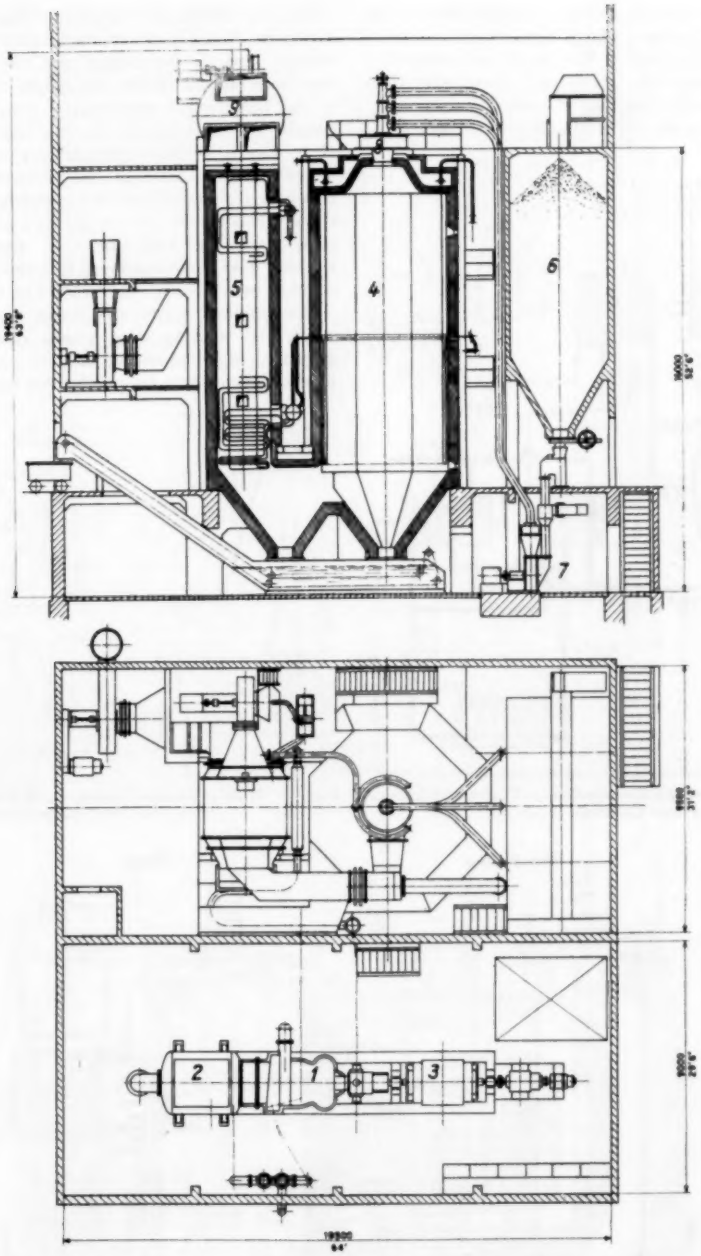


FIG. 22 LAYOUT OF MODERN 6600-KW CLOSED-CYCLE PLANT WITH PULVERIZED-COAL-FIRING

1 Tuco set
2 Precooler
3 Generator
4 Air-heater radiation section
5 Air-heater convection section
6 Coal bunker
7 Coal mill
8 Burner
9 Combustion air preheater

clear power plants up to 60 mw have been published recently by the AEC (10). Helium as the working medium requires more complex and multistage machinery. We actually are engaged in studying new design prospects for plants using gases other than air. Gas mixtures promise favorable solutions to machinery simplification because they lead to a reduction of turbomachine stages.

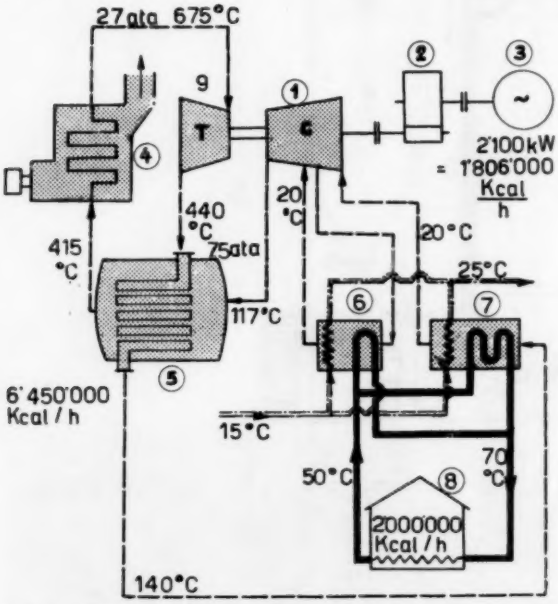


FIG. 21 SCHEME OF COMBINED CLOSED-CYCLE PLANT FOR POWER PRODUCTION AND CENTRAL HEATING

American Turbine Corporation (4) and Escher Wyss have been developing designs particularly adapted to United States requirements. Both fossil-fuel-fired and nuclear-reactor heat sources have been considered but the simplicity and the high-efficiency of the closed-cycle power-plant system make this plant especially desirable for the nuclear application. Accordingly, a 10 to 15-mw set has been designed which meets the requirements for a wide range of possible applications. The cross section of a turbomachinery set is shown in Fig. 27. Axial machinery was chosen for our plants above 10 mw because of its compactness and high efficiency. Pure nitrogen is used as a working fluid when employed in a nuclear plant; obviously air could be used when heat is supplied by fossil fuels.

One proposal employs a single, closed, nitrogen circuit, in which the working fluid passes directly through both the graphite-moderated reactor and the gas-turbine. A chemically fired air heater is to be installed for initial operation of the tur-

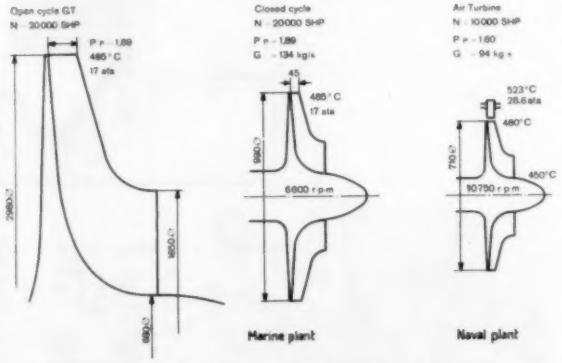


FIG. 24 RADIAL INFLOW TURBINE FOR ELEVATED-PRESSURE LEVELS AS USED IN CLOSED-CYCLE PLANT

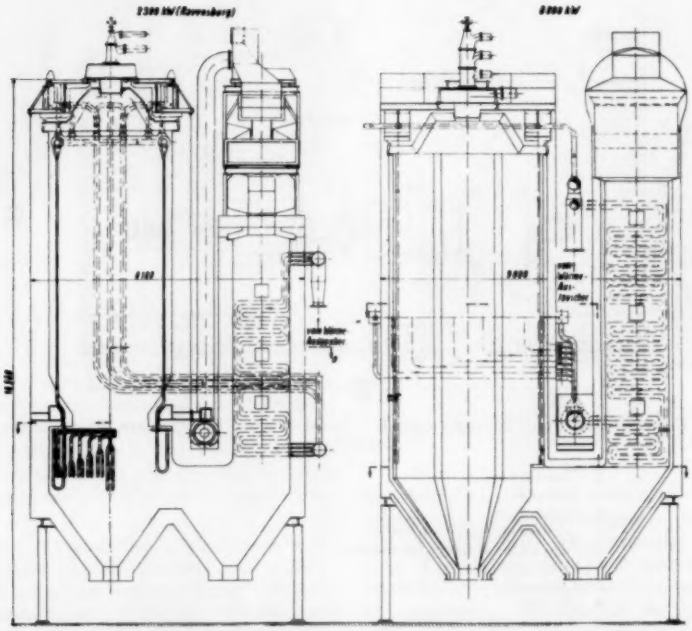


FIG. 23 COMPARISON OF 2000 AND 6600-KW COAL-FIRED AIR HEATERS

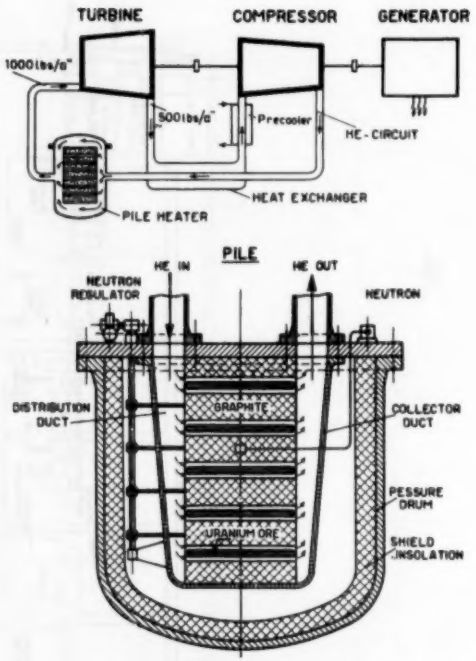


FIG. 25 (*above*) SCHEME OF A GAS-CYCLE REACTOR COMBINED WITH CLOSED-CYCLE GAS TURBINE IN ONE LOOP

(*Below*) 1945 PROPOSAL OF PRESSURE GAS-COOLED GRAPHITE-URANIUM REACTOR

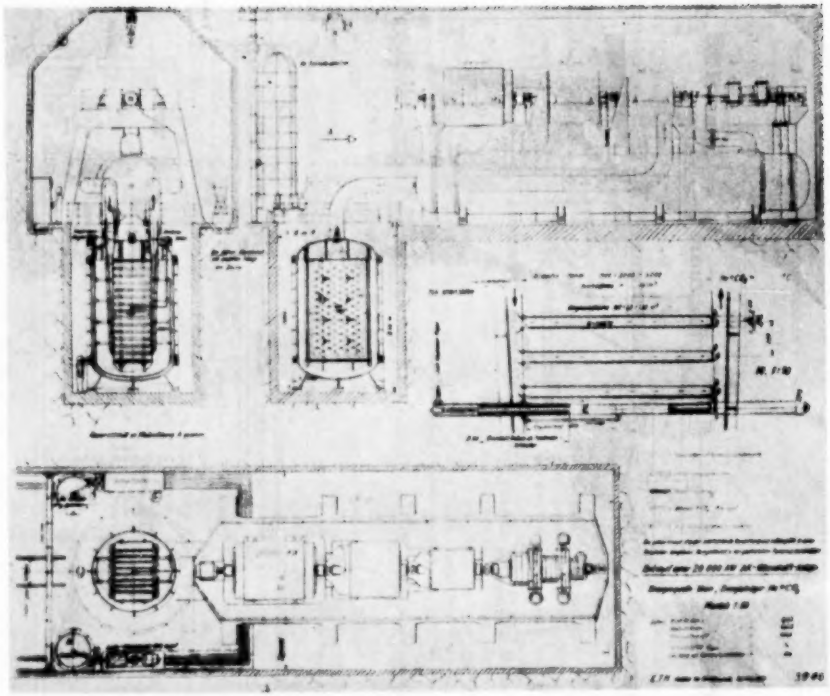


FIG. 26 EARLY PROPOSAL OF NUCLEAR POWER PLANT 20 MW WITH GAS-COOLED REACTOR AND CLOSED-CYCLE GAS TURBINE (1945)

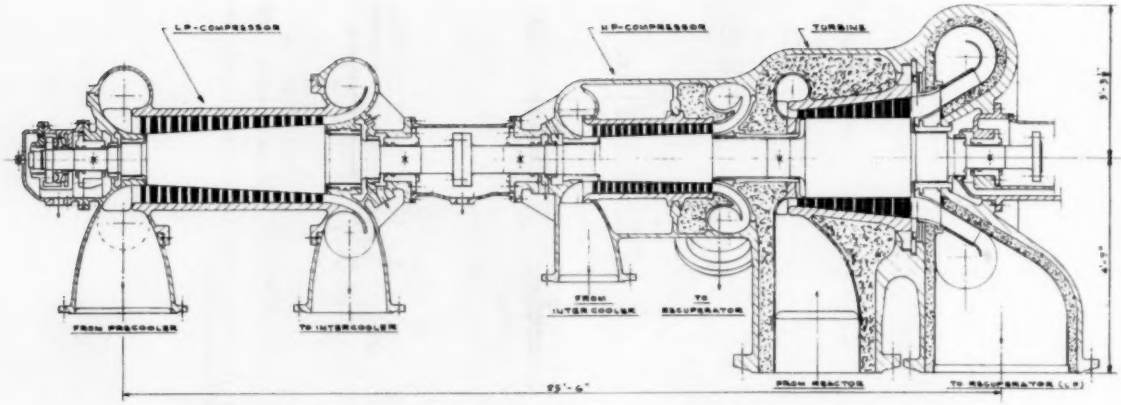
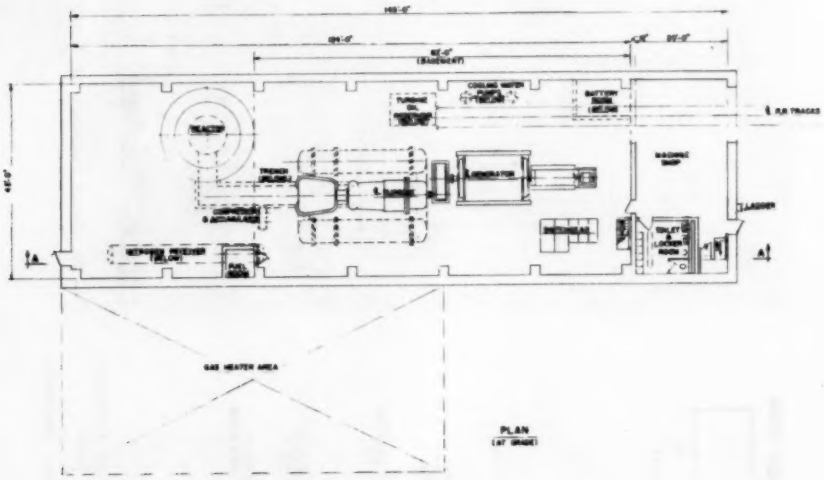


FIG. 27 CROSS SECTION OF 15–20-MW CLOSED-CYCLE MACHINE SET FOR USE WITH NITROGEN OR AIR



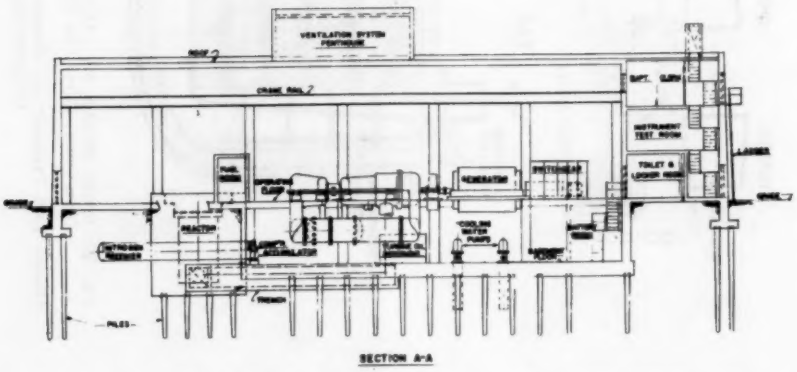


FIG. 28 LAYOUT OF 15-MW CLOSED-CYCLE GAS-TURBINE NUCLEAR POWER PLANT

bine before the reactor is placed in service. This heater will serve as a standby source of heat for the nuclear plant during periods of reactor experimentation or maintenance. Machinery (Fig. 27) is similar to that of Fig. 8 but utilizes a separate intercooler rather than an integral one. A typical layout incorporating this machinery is shown in Fig. 28. The simplicity and compact design of this 15-mw plant having over 30 per cent efficiency (at generator terminals) should be noted. The useful output and the work of compression are produced in a single turbine of 2 ft 6 in. OD, where nitrogen expands from 503 psia (1290 F) to 130 psia (325 F).

The use of high-pressure gas having good thermodynamic and nuclear properties results in a reactor of relatively small dimensions with small pressure losses.

BIBLIOGRAPHY

1 "The Escher Wyss—AK Closed-Cycle Turbine, Its Actual Development and Future Prospects," by Curt Keller, Trans. ASME, vol. 68, 1946, pp. 791–822.

2 "Closed-Cycle Gas Turbine, Escher Wyss—AK Development 1945–1950," by Curt Keller, Trans. ASME, vol. 72, 1950, pp. 835–850.

3 "The Closed-Cycle Gas Turbine Power Plant," by S. T. Robinson, ASME Paper No. 52-A-137.

4 List of Licensees—end of 1956:

Great Britain—John Brown & Company (Clydebank) Ltd., Clydebank.

United States—American Turbine Corporation, New York, (Licensors of Escher Wyss in U. S. A.).
Westinghouse Electric Corporation, East Pittsburgh (Pa.).
Nordberg Manufacturing Company, Milwaukee, Wis.

Germany —AEG Allgemeine Elektricitaets-Gesellschaft, Berlin.
GHH Gutehoffnungschütte Sterkrade Aktiengesellschaft, Werk Sterkrade, Oberhausen-Sterkrade.
Friedrich Krupp, Aktiengesellschaft, Essen.

Japan —Mitsubishi Zosen K.K., Tokyo.
Mitsui Shipbuilding & Engineering Co. Ltd., Tokyo.
Fuji Denki Seizo K.K., Tokyo.

5 "Conte rendu des essais officielles," by H. Quiby, Revue Politechnique Suisse, vol. 125, no. 23/24, June, 1945; also, *Oil Engine*, vol. 13, November, 1945, pp. 184–191.

6 "Einige Probleme der warmfesten hitzebeständigen Stähle vom Standpunkt des Verbrauchers," by W. Stauffer, *Schweiz. Archiv für angewandte Wissenschaft und Technik*, heft 12, vol. 17, 1951, pp. 353–364.

"The Scaling Behavior of High-Strength Heat Resisting Steels in Air and Combustion Gases," by W. Stauffer and H. Kleiber, *Journal of the Iron and Steel Institute*, vol. 156, 1947, pp. 181–188.

7 "Closed-Cycle Air Turbine Installation for Marine Propulsion," by Curt Keller and W. Spillmann, *Oil Engine and Gas Turbine*, vol. 21, 1953, pp. 317–319; January, 1954, pp. 355–358. Also, "Beispiele geschlossener Heissluft-Turbinenanlagen für Kriegs- und Handelsschiffe," Jahrbuch der Schiffbautechn. Gesellschaft, Hamburg, Germany, Bd. 48, 1954.

8 "Applied Atomic Energy," by Tom Sawyer, Prentice-Hall, Inc., New York, N. Y., 1946, p. 141.

9 "Nuclear Reactors for Large Gas Turbines," by T. Jarvis, and "The Closed-Cycle Nuclear Gas Turbine Power Plant," by S. T. Robinson, ASME Symposium on Nuclear Gas Turbines, Washington, D. C., 1956.

10 "Design Study 60 MW Closed-Cycle Gas Turbine Nuclear Power Plant," by S. T. Robinson, American Turbine Corporation, published by AEC, December, 1954.

Research on Application of Cooling to Gas Turbines

By J. B. ESGAR,[1] J. N. B. LIVINGOOD,[2] and R. O. HICKEL,[3] CLEVELAND, OHIO

The use of turbine cooling in gas-turbine engines can offer many performance benefits but, at the same time, it may result in added complication to the engine. The advantages that turbine cooling can offer to the engine designer and the results of some of the research that has been expended on the cooling of gas-turbine engines are discussed.

Introduction

THE application of turbine cooling to gas-turbine-type aircraft engines permits increases in turbine-inlet temperature to the point where engine power can be increased greatly. For some applications the specific fuel consumption of the engine also can be improved. Furthermore, the use of cooling will permit increased allowable stress levels. As a result, it is possible to increase the mass-flow rate through the turbine and, in all probability, the turbine can be made more reliable. On the other hand, the use of cooling introduces more complications into the engine design and, for some modes of flight operation, high turbine-inlet temperature causes poor fuel economy. Because it is necessary to weigh the advantages and disadvantages of turbine cooling as well as to consider many modes of flight and engine operation, there is often considerable disagreement among engine designers as to whether turbine cooling is really worth while.

The latest discussion presented to this Society on turbine cooling by a member of the NACA staff was by O. W. Schey in 1948 (1).[4] The present paper will try to show, within security limitations, when it is desirable to utilize cooling in gas-turbine engines and will present some of the results obtained at the Lewis Flight Propulsion Laboratory from research directed towards the application of cooling to gas turbines since the Schey paper (1) was published.

Benefits From Turbine Cooling

Cooling of the turbine blades and disks by means of either air or liquid results in degrees of freedom in engine design and operation not possible with conventional types of engines. Some of the things made possible through use of turbine cooling are as follows:

Increased Turbine-Inlet Temperature. The increase in turbine-inlet temperature made possible by cooling the turbine causes higher tail-pipe jet velocities and thus greater thrust for turbojet engines and increased turbine-shaft power for engines such as the turboprop. For either type of engine, the power per unit of engine weight or per pound of air flow can be increased substantially. This can be seen in Fig. 1. These performance curves were calculated by the methods of (2) for a flight Mach number of 0.9 and an altitude of 40,000 ft. A flight Mach number of 0.9 is low for many turbojet-engine applications and is somewhat high for present turboprop-engine applications. This intermediate flight Mach-number value was selected, however, so that the effects of turbine-inlet temperature on engine performance could be made at a single flight condition that is reasonable for each engine type.

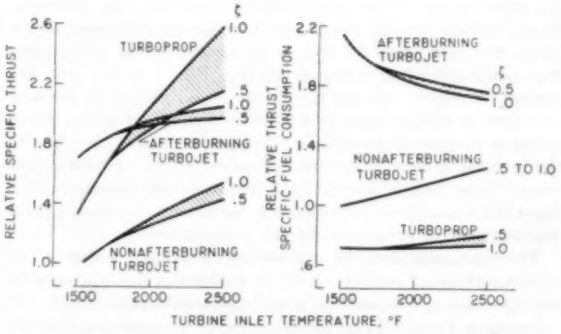


Fig. 1 Effect of Turbine-Inlet Temperature on Relative Specific Thrust and Relative Thrust Specific-Fuel Consumption. Effects of Air Cooling Included. Flight Mach Number, 0.9; Altitude, 40,000 Ft

The performance shown is relative to that of a nonafterburning turbojet engine with a 1540 F turbine-inlet temperature. All engines were assumed to have a compressor-pressure ratio of 12 and compressor and turbine adiabatic efficiencies of 0.85. For the afterburning turbojet engine, the gas temperature at the tail-pipe nozzle was assumed to be 3000 F. In order to compare all engines on the same basis, the shaft power of the turboprop engine was converted to thrust by assuming that a propeller thrust was added to that of the jet thrust. The effects of air-cooling the turbine stator and rotor blades to a metal temperature of 1240 F are included in the performance results shown. The cooling air was assumed to be bled from the discharge of the compressor. After cooling the blades, the air was considered to be mixed with the exhaust gases downstream of the turbine. A blade-cooling effectiveness parameter ζ is used to show air-cooling effects on performance. This parameter is defined as

$$\zeta = \frac{T_{a,\,out} - T_{a,\,in}}{T_B - T_{a,\,in}} \quad \ldots\ldots\ldots\ldots\ldots\ldots [1]$$

where

$T_{a,\,in}$ = cooling-air inlet temperature at blade base
$T_{a,\,out}$ = cooling-air outlet temperature at blade tip
T_B = average blade temperature

The blade-outside heat-transfer coefficients required to obtain heat-rejection rates were found by the methods of (3). The value

---

[1] Associate Chief, Turbine Cooling Branch, NACA Lewis Flight Propulsion Laboratory.

[2] Head, Section A, Turbine Cooling Branch, NACA Lewis Flight Propulsion Laboratory.

[3] Head, Section B, Turbine Cooling Branch, NACA Lewis Flight Propulsion Laboratory.

[4] Numbers in parentheses refer to the Bibliography at the end of the paper.

Contributed by the Gas Turbine Power Division and presented at the Semi-Annual Meeting, Cleveland, Ohio, June 17–21, 1956, of The American Society of Mechanical Engineers.

Note: Statements and opinions advanced in papers are to be understood as individual expressions of their authors and not those of the Society. Manuscript received at ASME Headquarters, May 4, 1956. Paper No. 56—SA-54.

of ζ for the better air-cooled turbine blades approaches or exceeds a value of 1.0. (The reason a value of 1.0 can be exceeded is that the blade temperature near the blade tip is higher than the average blade temperature.) Turbine blades that would have values of $\zeta = 0.5$ would be of a poor cooling design.

From Fig. 1 it can be seen that the thrust of turboprop and nonafterburning turbojet engines can be increased greatly by increasing turbine-inlet temperatures. For the turboprop engine there is only a slight increase in specific fuel consumption. The specific fuel consumption for nonafterburning turbojet engines increases more with turbine-inlet temperature, but it is still more economical than using an afterburner. Afterburning to 3000 F results in the highest thrust by far for the three types of engines at low turbine-inlet temperatures, but the specific fuel consumption is very high. The use of higher turbine-inlet temperatures reduces the specific fuel consumption of afterburning engines and at the same time results in increased thrust.

From a performance point of view, therefore, it appears that there are no disadvantages to increasing turbine-inlet temperatures for turboprop and afterburning turbojet engines. The cooling effectiveness of the turbine blades should be high, however, for the turboprop to get the greatest increases in power and, at the same time, not cause the specific fuel consumption to increase. For nonafterburning turbojet engines, Fig. 1 shows that increases in relative specific thrust are accompanied by marked increases in relative specific fuel consumption. The rate of increase in relative specific thrust, however, is about twice that of the increase in relative specific fuel consumption. As will be discussed later, higher turbine-inlet temperatures at supersonic-flight Mach numbers can be desirable for nonafterburning engines when considered along with aircraft performance.

The same general trends as those shown in Fig. 1 would be obtained with liquid cooling. A study of effects of liquid cooling on the efficiency of a turboprop-type engine is presented in (4).

Increased Turbine Stress Levels. Since engine power is directly proportional to the gas-flow rate, it is desirable to increase the turbine-flow capacity by use of longer turbine blades. Longer blades, in turn, result in higher turbine-blade stresses. The relationship between allowable blade stress and blade temperature in Fig. 2 shows the 100-hr stress-rupture properties for three alloys.

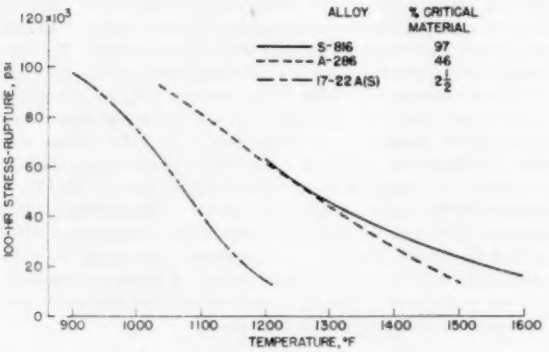


FIG. 2 STRESS-RUPTURE PROPERTIES FOR POSSIBLE AIR-COOLED TURBINE-BLADE MATERIALS

The upper levels of the curves are terminated where stress-rupture properties no longer determine the permissible stress. The alloy S-816 is commonly used in present gas-turbine engines. At a temperature of 1500 F (about standard blade temperature for present engines), the maximum allowable stress is 24,000 psi. If, however, the temperature is reduced only 100 deg F by cooling, the allowable stress can be increased by about 35 per cent, with further increases obtainable at lower temperatures. Below temperatures of about 1200 F, however, other materials such as A-286 possess better strength properties, with the possibility of operating at stresses over 90,000 psi—over $3\frac{1}{2}$ times the allowable stress for present engines. As a result, turbine-gas flow capacity can be increased greatly. Turbine work capacity also can be increased by operating at higher wheel speeds that are possible with increased allowable stresses. Further reduction in temperature makes possible the utilization of high-strength steels such as Timken 17-22A(S). This type of material offers only slight increases in possible operating stress over A-286, but the critical material content of 17-22A(S) is eliminated almost completely.

Increased Turbine Reliability. The design of turbine blades is unique in machine or structural design in that little or no factor of safety is employed. The blades are designed on a basis of stress-rupture properties for some specified life, which may be only from 100 to 300 hr with a design safety factor of 1.0. This low safety factor occurs because generally it would be desirable to operate engines at higher turbine-inlet temperatures than is possible without cooling to improve engine power. As a consequence, it has been necessary to operate uncooled engines at temperatures as high as possible, with a resultant sacrifice, to a considerable extent, in reliability. Cooling provides a means of increasing the factor of safety in turbine blades without lowering turbine-inlet temperatures and sacrificing engine performance. Fig. 3 shows the stress-rupture properties of A-286 for lives of 100 and 1000 hr. A tenfold increase in life can be obtained for blades made of A-286 by decreasing the blade temperature from 50 to 90 deg F (depending on the stress level) through use of turbine cooling. By such design methods, engine life and reliability could be increased substantially.

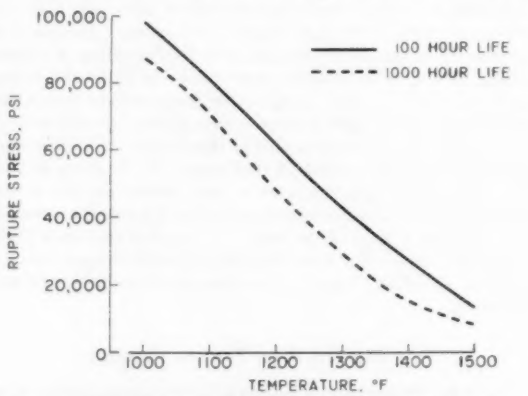


FIG. 3 STRESS-RUPTURE PROPERTIES OF A-286 ALLOY

A word of caution should be injected at this point, however. Some cooled-turbine research has been conducted which indicates that considerable potential for increased reliability exists, but much more work is required to verify this in extended actual engine and flight operation. Cooled turbine blades are constructed in a manner completely different from those blades that operate in current engines. As a result, considerably more blade development will be required before definite statements regarding improvements in engine reliability can be made.

Improved Aircraft Performance. One method of evaluating aircraft performance is by the Breguet range equation which can be written

$$R \propto \eta \frac{L}{D} \log_e \frac{W_g}{W_g - W_f} \quad \ldots\ldots\ldots\ldots\ldots\ldots [2]$$

Equation [2] can also be written as

$$R \propto \eta \frac{L}{D} \log_e \frac{1}{\frac{W_s}{W_g} + \frac{W_p}{W_g} + \frac{W_e/F}{L/D}} \quad \ldots\ldots [3]$$

where

R = range
η = over-all engine efficiency
L/D = lift-drag ratio
W_g = initial airplane gross weight
W_f = initial fuel weight
W_p = pay-load weight
W_s = structural weight
W_e/F = thrust specific engine weight

So far as the engine is concerned, two terms in Equation [3] affect aircraft range; namely, the over-all efficiency η and the thrust specific engine weight W_e/F. At a given flight speed the over-all efficiency is inversely proportional to the thrust specific fuel consumption. In general the over-all efficiency increases with increasing flight speed. The aircraft range improves with increasing over-all efficiency and decreasing thrust specific engine weight.

The use of turbine cooling to increase stresses so that the mass-flow capacity of the engine can be increased will probably result in reduced specific engine weight. This would result in a beneficial effect on range and could be obtained without increasing turbine-inlet temperature.

From Fig. 1 it can be seen that for a nonafterburning turbojet engine the over-all engine efficiency will decrease as turbine-inlet temperature is increased, but at the same time the thrust specific engine weight also decreases because of the large increases in thrust. There is a counterbalancing effect since both the efficiency and the specific engine weight decrease simultaneously in nonafterburning engines. Which effect will finally predominate results in a lengthy study and will vary with different types of airplanes and engine designs. Without going into the details, it can be stated that at high supersonic-flight speeds and at very high altitudes, high turbine-inlet temperatures result in increased aircraft range, because at these conditions the reduction in specific engine weight more than outweighs the effect of decreased efficiency.

It is obvious from Fig. 1 and Equation [3] that aircraft performance where turboprop or afterburning turbojet engines are used can be improved by increasing turbine-inlet temperatures. Performance gains, however, for afterburning engines are marginal for turbine-inlet temperatures above about 2000 F. At higher flight Mach numbers slightly higher turbine-inlet temperatures are desirable.

Design Considerations for Cooled Turbine Engines

In order to capitalize to the fullest extent on the performance benefits possible through use of turbine cooling the basic engine design should be different from the conventional practice in uncooled engines. This means therefore that adapting uncooled engines so that the turbines can be cooled and the turbine-inlet temperature can be raised will not produce engines of the high performance possible if cooling is considered in the original concept of the engine design.

Turbine-stress limitations encountered in uncooled engines are not valid for cooled engines. Because of much higher allowable stresses the turbines can be rotated faster and the blades made longer to permit use of higher turbine power and higher flow capacity for a given engine size. The compressor flow, size, and rotational speed are in turn affected by these possible changes in the turbine. By proper design the total engine power and specific engine weight can be improved to a greater degree, by building in all the advantages possible through use of turbine cooling, than is possible by increasing turbine-inlet temperature alone.

Application of Cooling to Turbines

From the preceding discussion, it can be seen that it would often be quite desirable to use turbine cooling to permit higher turbine-inlet temperatures or high stresses for both turbojet and turboprop engines. An evaluation of several means of turbine cooling will be discussed next to show how the turbine can be made to withstand these conditions.

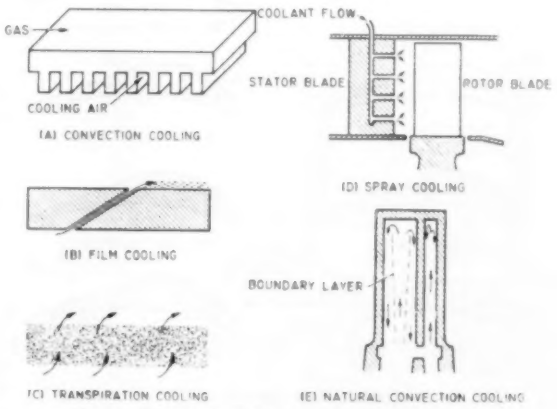


Fig. 4 Examples of Several Cooling Methods

Methods of Blade Cooling. Fig. 4 shows three methods of air cooling and two methods of liquid cooling. The most conventional air-cooling method used in heat-transfer processes is convection-cooling, Fig. 4(*A*). With this method, it is desirable to increase the heat-transfer-surface area on the heat-rejection side of the apparatus. The fins shown in Fig. 4(*A*) serve this purpose. This method of cooling has been used successfully on air-cooled piston engines for many years. Film cooling is illustrated in Fig. 4(*B*). A film of cool air is introduced through slots to form an insulating layer between the hot gases and the cooled surface. The thermal conductivity of air is very low so that it is a good insulation medium, but the effectiveness of the layer of air is lost some distance downstream of the slot because of the mixing of the coolant with the hot gases. This disadvantage is eliminated by transpiration cooling, Fig. 4(*C*), because air is passed continuously through the entire area of a porous surface. Transpiration cooling is the most effective method of air-cooling known at the present time. A comparison of the effectiveness of these three methods of cooling is given in (5).

Fig 4(*D*) shows a form of liquid cooling where the liquid is sprayed into the gas stream so that it will impinge upon and cool the rotor blades. With this cooling device the liquid used for the coolant is lost. A closed-type liquid-cooling system where the coolant is recirculated also can be employed. This system is similar to that used in automobiles. Several methods can be used that will result in coolant circulation within the blades without special coolant pumps. A natural-convection cooling system is illustrated in Fig. 4(*E*). The natural-convection circulation can be compared to that in a home hot-water tank where the heated water rises to the top of the tank and the cooler water sinks to the bottom. With the high centrifugal-force fields that are set up in turbine rotors, the circulation rates due to natural convection becomes very high. The liquid coolant can be circulated within a

single blind hole. In such a system, relatively cool liquid is forced radially outward through the central portion of the hole by centrifugal force. As the coolant near the wall is heated, its density becomes less than that of the cool liquid core and it flows radially inward. In a system where crossover holes are used, the coolant would flow out one hole, cross over at the tip, and flow inward in the next hole.

Air-Cooled Turbine Blades. Air-cooled turbine blades of about 2-in. chord utilizing the air-cooling methods shown in Figs. 4(*A*) to (*C*) are illustrated in Fig. 5. The hollow blade was used by the

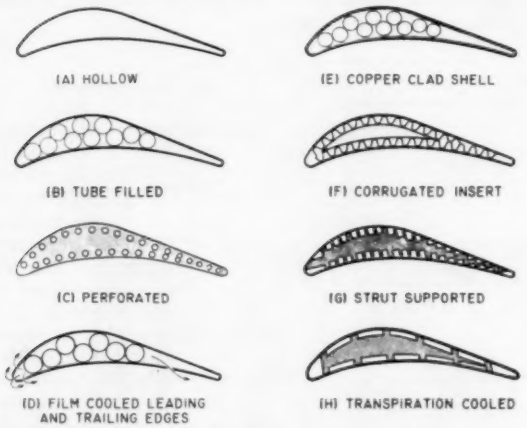


FIG. 5 AIR-COOLED TURBINE-BLADE CONFIGURATIONS

Germans in some of their engines in 1945. A survey of their work on cooling turbojet engines and turbosuperchargers is given in (6) to (8). The cooling effectiveness of the hollow blade is so low that excessive quantities of cooling air are required; consequently, efforts were made to provide added internal heat-transfer surface area. The tube-filled blade, Fig. 5(*B*), was an early attempt of the NACA to provide this extra surface area. Some results of tests on and methods of manufacturing this type of blade are given in (9) to (13). Although more recent blade developments have led to superior blade configurations, much valuable information has been obtained from the tube-filled configuration.

The British (14) have used a somewhat different method of approach to the problem of adding internal surface area. Instead of packing a hollow shell with tubes and brazing the assembly together, holes were provided near the periphery of solid blades, Fig. 5(*C*), by drilling or other methods. Herein this type of blade will be called a perforated blade.

Cooling of the leading and trailing edges is often difficult with tube-filled blades. In an attempt to improve cooling effectiveness in these regions, film cooling, Fig. 5(*D*), and copper-clad shells, Fig. 5(*E*), were investigated. The copper-clad blade is a type of structure similar to that of copper-clad kitchen utensils in which the copper spreads the heat over the entire area of the utensil. In the case of cooled turbine blades, the copper cladding is on the inner surface of the blade shell. Results of heat-transfer tests on film-cooled and copper-clad blades are reported in (15) to (17).

The effects on reduction of temperatures of turbine-blade leading and trailing edges are shown by the experimental data in Fig. 6. It will be noted that the blade-temperature gradients can be reduced by use of film-cooled and copper-clad blades, but the leading and trailing-edge cooling is done at the expense of increased temperatures in the midchord region of the blades. The film-cooled blade shown operated only slightly cooler at the leading edge than did the 10-tube blade. The amount of cooling obtained in this region is a function of the configuration and the relative pressures on the inside and outside of the blade. At higher flow rates this cooling becomes more effective because a greater portion of the air is bled from the slots. The blade must be designed specifically for the application for which it is intended and the coolant supply pressure must be higher than the stagnation pressure at the blade leading edge. More effective leading-edge cooling has been found in other film-cooled blades. Durability of this type of blade was found to be a serious problem as reported in (12). Research was conducted in Germany on blades having film cooling around the complete periphery (18), and similar blades have been built in this country. Although cooling is adequate for some cases, durability of rotor blades is usually poor. This type of blade probably would be more successful when applied to stators, provided that cooling-air pressure is sufficiently high.

Although the cooling of the copper-clad blade looks better than for the film-cooled blade in Fig. 6, the weight increase due

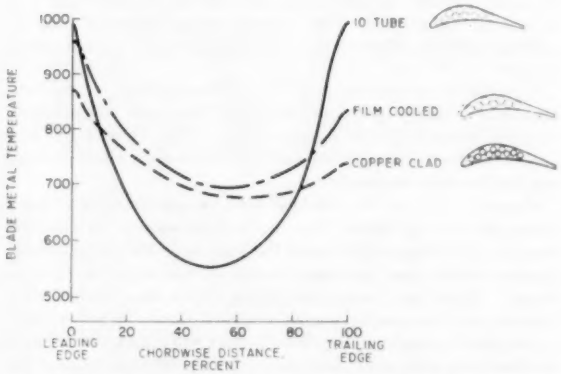


FIG. 6 EFFECT OF SPECIAL METHODS OF COOLING LEADING AND TRAILING EDGES ON METAL TEMPERATURE

(Engine speed, 10,000 rpm; effective gas temperature, 1070 F; cooling-air temperature, 130 F; ratio of cooling-air flow to combustion-gas flow ratio, 0.045.)

to the addition of copper raises the stress for rotor blades so much that the gains from cooling are practically eliminated. In addition, at the blade temperatures required in engine operation, copper oxidation is rapid and severe.

A better method of reducing chordwise temperature gradients for rotor blades is to utilize thinner blade shells. The leading and trailing edges would then have larger cooling-air passages which would insure an adequate supply of cooling air in those regions. Configurations that can extend the augmented heat-transfer surface well into the leading and trailing edges also should be used. Such a blade, Fig. 5(*F*), is one with corrugated fins (11). An island is usually provided in the middle of the coolant passage so that the corrugations can be of uniform amplitude. The core of the island is blocked off from the cooling air.

In all the turbine blades discussed up to this point, with the possible exception of the perforated blade, the blade shell has been the primary member for carrying the stresses due to centrifugal forces. Since the shell is exposed to the gas stream, it is also the hottest member of the blade; therefore its stress-carrying capacity is lower than that of cooler portions of the blade. For this reason it seems practical to design blades where the main stress-carrying member, or strut, is submerged inside the coolant passage (19), where it will be at a lower temperature than the blade shell, Fig. 5(*G*). This type of construction can be used with

either an impermeable or a permeable shell, in which case the latter blade is called transpiration-cooled, Fig. 5(*H*). More of the details of this type of construction are illustrated in Fig. 7. The porous shell could be made from several materials, the most probable being woven-wire cloth (20) to (22) or porous sintered materials made from powdered metal (23).

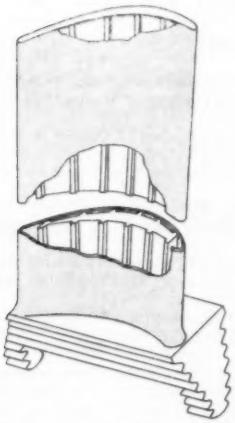


FIG. 7 TRANSPIRATION-COOLED, STRUT-SUPPORTED BLADE

The design and manufacture of transpiration-cooled blades are somewhat difficult because the coolant-flow rate is a function of permeability of the shell and the pressures inside and outside of the blade. Because of the high-pressure gradients on the outside of the shell, large variations in shell permeability are required. Some of the advantages and problems of transpiration cooling are discussed in (24), and methods of accounting for rotational effects on cooling-air distribution are given in (25).

Many analytical studies have been made on methods of calculating cooling-air requirements for the various types of blades shown in Fig. 5. Calculated coolant flows for five of the blades for a range of turbine-inlet gas temperatures are shown in Fig. 8. These flows are on a relative basis because the absolute magnitude of the flows is dependent upon many conditions, such as the type of engine, size of blade, flight speed, and flight altitude. A review of blade-temperature calculation methods is given in (26). More specifically, the coolant flows for the hollow, tube-filled, and corrugated-insert blade were calculated by methods presented in (27) to (29), the strut blade by (19) and the porous blade by (24), (30), and (31). The load-carrying member of all blades was assumed to be cooled to 1240 F.

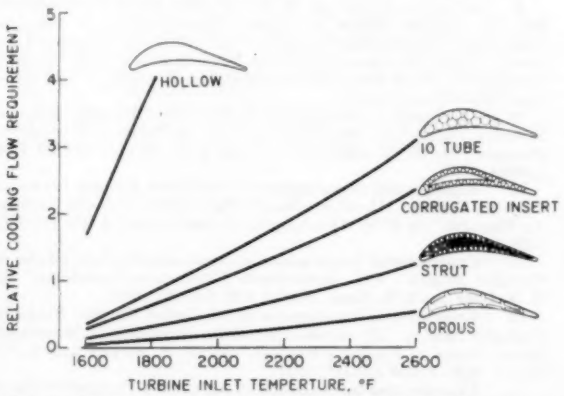


FIG. 8 COMPARISON OF RELATIVE COOLANT-FLOW REQUIREMENTS FOR SEVERAL TYPES OF AIR-COOLED TURBINE BLADES

The use of plain hollow blades is impractical for cooled turbines because the coolant-flow requirements are exorbitant. On the other extreme is the transpiration-cooled turbine blade which ideally requires only very small amounts of coolant even for large increases in turbine-inlet temperature. Generally, it is desirable to use blades requiring the smallest coolant flow, but other factors such as fabrication problems, cooling-air pressure required, clogging (such may be encountered with transpiration-cooling), blade weight, and durability also must be considered. As a result, the final choice of the blade design is up to the designers' discretion and will be dependent upon the over-all design and proposed application of the engine.

Air-Cooled Turbine-Disk Configurations. The use of air-cooled turbine blades will require a type of turbine-rotor construction different from that in current use. There is, however, a considerable amount of freedom in the type of design possible. Two main types of turbine rotor are the split disk, Fig. 9, and the shrouded disk, Fig. 10. A discussion of these disk configurations is given in (32). With either type of construction the cooling air can be supplied from the upstream direction, the downstream direction, or through a hollow turbine shaft. With any of these possible types of construction, internal vanes are required in the turbine rotor to direct and help pump the cooling air out to the blades. Up to the present time, experimental tests have been conducted on several turbines with the type of disk configuration shown in Fig. 9(*b*), and some of the results presented in (32) to (34) indicate that disk cooling will be adequate with the amount of air required for blade cooling.

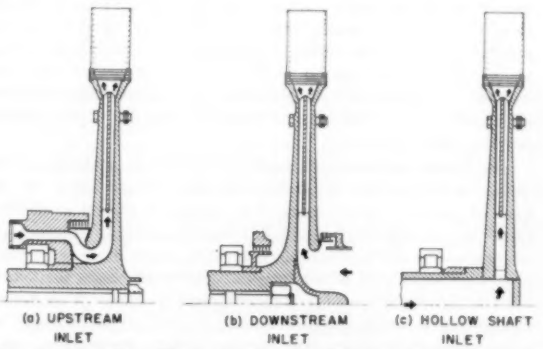


FIG. 9 SPLIT-DISK-TYPE, AIR-COOLED, TURBINE CONFIGURATIONS

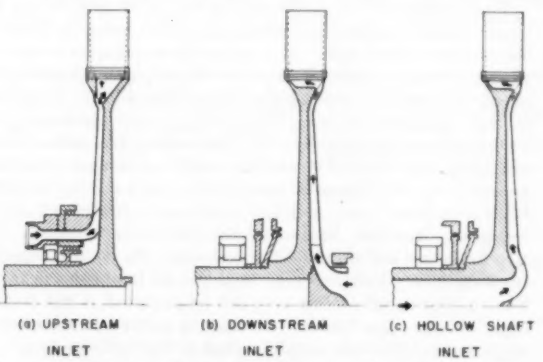


FIG. 10 SHROUDED-TYPE, AIR-COOLED, TURBINE-DISK CONFIGURATIONS

Liquid Cooling. Internal water-cooled turbines have been operated in this country and in Europe (35) to (37). This method of operation was logical because of the excellent heat-transfer characteristics of water. Water as a turbine coolant, however, has one very serious disadvantage—the boiling point is so low that unless the entire coolant system is under very high pressurization the turbine is overcooled and the heat-rejec-

tion rates are excessive. A further disadvantage occurs at high flight speeds because the ram-air temperature exceeds the boiling temperature of water at normal pressures, and heat rejection in a ram-air heat exchanger becomes difficult. The ram-air temperature reaches 212 F at a flight Mach number of about 1.2 at standard sea-level conditions and at a Mach number of about 1.9 in the stratosphere.

Pressurization of the entire coolant system would offer relief, but in order to pressurize the system to a point where water could be at a temperature that would not result in overcooling of the turbine, the system would have to be pressurized to approximately 3000 psi. A method of eliminating this difficulty would be to use special coolants such as Dowtherm or liquid metals which have higher boiling points than water at normal pressure levels. The design of liquid-coolant systems requires a knowledge of heat-transfer coefficients for forced and free convection. Considerable experimental and analytical heat-transfer data (38) to (41), have been obtained that are applicable to liquid-cooled turbines.

The British are giving thought to the use of a thermosiphon liquid-coolant system for turbine blades (42) and (43). In this type of system, a small amount of coolant is placed in the turbine—not enough to fill the blade-coolant passages. The coolant evaporates in the turbine blades. The vapor is passed over a heat exchanger and is condensed. The condensate then flows back to the turbine blades to be evaporated again. This system holds promise for some applications. Its primary advantage over other types of liquid-coolant systems is that the internal pressures in the blades can be considerably lower.

Water-spray cooling as discussed in (44) offers possibilities for certain limited applications where a boost in power is required for short periods of time. This method of running for prolonged periods of time results in excessive liquid consumption in the engine because the water used as the coolant is lost. Using this method of cooling the rotor blades alone permits increasing the turbine-inlet temperature to the point where the stator blades provide a limitation. For relatively small rotor blades, the temperatures can be reduced adequately by spraying water onto the blade from a few (varying from 1 to 8) stationary circumferential locations. However, because of the low water velocities relative to the gas and turbine-wheel velocities, the coolant supplied from stationary locations only impinges on the blades near the leading edges, and cooling may not be provided at the trailing edges of large rotor blades. A "rotating gutter" on the rotor as suggested in (44) could supply coolant to the trailing edges.

In order to provide spray cooling to the stator blades so that higher turbine-inlet temperatures could be obtained than are possible with rotor cooling alone would require a more complicated water-spray system, and, of course, the coolant flow would have to be increased. The total flow would be about 2 to 4 times the weight-flow rate of fuel required.

Consideration has been given to fuel injection ahead of the turbine for afterburner use (45). Although very few experimental temperature data are available, it is not believed that the quantities of fuel required for suitable afterburning temperatures would be adequate for use in cooling of the blades; water would still have to be injected to augment the cooling from the fuel spray.

The time that water-spray cooling would probably be most beneficial would be during take-off of aircraft having marginal take-off power. For such a case, the turbine-inlet temperature could be raised a few hundred degrees. The water that would be used for cooling would be expended during the take-off run, and the weight of the coolant would not have to be carried during the remainder of the flight. This method of cooling would also be useful for short bursts of emergency power during flight.

In conclusion it can be stated that, for many applications of gas-turbine engines, turbine cooling can offer substantial performance benefits. For these cases there are many methods of turbine cooling that are possible, leaving considerable choice to the engine designer.

BIBLIOGRAPHY

1 "The Advantages of High Inlet Temperature for Gas Turbines and Effectiveness of Various Methods of Cooling of Blades," by O. W. Schey, ASME Paper No. 48—A-105.

2 "Methods for Rapid Graphical Evaluation of Cooled or Uncooled Turbojet and Turboprop Engine or Component Performance (Effects of Variable Specific Heat Included)," by J. B. Esgar and R. R. Ziemer, NACA TN 3335, 1955.

3 "Extension of Boundary-Layer Heat-Transfer Theory to Cooled Turbine Blades," by W. B. Brown and P. L. Donoughe, NACA RM E50F02, 1950.

4 "Effect of Turbine-Blade Cooling on Efficiency of a Simple Gas Turbine Power Plant," by W. M. Rohsenow, Trans. ASME, vol. 78, 1956, pp. 1787–1794.

5 "Comparison of Effectiveness of Convection-, Transpiration-, and Film-Cooling Methods With Air as Coolant," by E. R. G. Eckert and J. N. B. Livingood, NACA Report 1182, 1954.

6 "A Survey of German Hollow Turbine Blade Development. Pt. 1—Initial Investigations and Developments," by E. N. Petrick, Purdue University, Purdue Research Foundation, published by USAF-AMC, Wright-Patterson Air Force Base, Dayton, Ohio, October, 1949. (USAF Contract W33-038-ac-17625.)

7 "A Survey of German Hollow Turbine Blade Development. Part II—The Design Features and the Production of Hollow Turbine Blades and Hollow Turbine Rotors for Turbosuperchargers and Turbojet Engines," by E. N. Petrick, Purdue University, Purdue Research Foundation, published by USAF-AMC, Wright-Patterson Air Force Base, Dayton, Ohio, December, 1949. (USAF Contract W33-038-ac-17625.)

8 "A Survey of German Hollow Turbine Blade Development. Part III—Turbine Blade Vibrations; Considerations on the Improvement of Hollow Blade Designs," by E. N. Petrick, Purdue University, Purdue Research Foundation, pub. by USAF-AMC, Wright-Patterson Air Force Base, Dayton, Ohio, March, 1950. (USAF Contract W33-038-ac-17625.)

9 "Experimental Investigation of Air-Cooled Turbine Blades in Turbojet Engine. I—Rotor Blades With 10 Tubes in Cooling-Air Passages," by H. H. Ellerbrock, Jr. and F. S. Stepka, NACA RM E50I04, 1950.

10 "Experimental Investigation of Air-Cooled Turbine Blades in Turbojet Engine. III—Rotor Blades With 34 Steel Tubes in Cooling-Air Passages," by R. O. Hickel and G. T. Smith, NACA RM E50J06, 1950.

11 "Experimental Investigation of Air-Cooled Turbine Blades in Turbojet Engine. VII—Rotor-Blade Fabrication Procedures," by R. A. Long and J. B. Esgar, NACA RM E51E23, 1951.

12 "Experimental Investigation of Air-Cooled Turbine Blades in Turbojet Engine. IX—Evaluation of the Durability of Noncritical Rotor Blades in Engine Operation," by F. S. Stepka and R. O. Hickel, NACA RM E51J10, 1951.

13 "Experimental Investigation of Air-Cooled Turbine Blades in Turbojet Engine. X—Endurance Evaluation of Several Tube-Filled Rotor Blades," by J. B. Esgar and J. L. Clure, NACA RM E52B13, 1952.

14 "An Experimental Single-Stage Air-Cooled Turbine. II—Research on the Performance of a Type of Internally-Air-Cooled Turbine Blade," by D. G. Ainley, *Aircraft Engineering*, vol. 25, September, 1953, pp. 269–276.

15 "Experimental Investigation of Air-Cooled Turbine Blades in Turbojet Engine. IV—Effects of Special Leading- and Trailing-Edge Modifications on Blade Temperature," by H. H. Ellerbrock, Jr., C. F. Zalabak, and G. T. Smith, NACA RM E51A19, 1951.

16 "Experimental Investigation of Air-Cooled Turbine Blades in Turbojet Engine. VI—Conduction and Film Cooling of Leading and Trailing Edges of Rotor Blades," by V. L. Arne and J. B. Esgar, NACA RM E51C29, 1951.

17 "Experimental Investigation of Air-Cooled Turbine Blades in Turbojet Engine. VIII—Rotor Blades with Capped Leading Edges," by G. T. Smith and R. O. Hickel, NACA RM E51H14, 1951.

18 "Temperature Measurement on Two Stationary Bucket Profiles for Gas Turbines with Boundary-Layer Cooling," by K. H. Kuepper, Trans. No. F-TS-1543-RE, Air Materiel Command, U. S. Air Force, January, 1948. (ATI No. 18576, CADO).

19 "Use of Electric Analogs for Calculation of Temperature of Cooled Turbine Blades," by H. H. Ellerbrock, Jr., E. F. Schum, and A. J. Nachtigall, NACA TN 3060, 1953.

20 "Wire Cloth as Porous Material for Transpiration-Cooled Walls," by E. R. G. Eckert, M. R. Kinsler, and R. P. Cochran, NACA RM E51H23, 1951.

21 "Experimental Investigation of Air-Flow Uniformity and Pressure Level on Wire Cloth for Transpiration-Cooling Applications," by P. L. Donoughe and R. A. McKinnon, NACA TN 3652, 1956.

22 "New Sintered-Wire Porous Metal Shows Promise for Aviation Use," by Irving Stone, *Aviation Week*, vol. 63, July 11, 1955, pp. 65–69.

23 "Experimental Investigation of Coolant-Flow Characteristics of a Sintered Porous Turbine Blade," by E. R. Bartoo, L. J. Schafer, Jr., and H. T. Richards, NACA RM E51K02, 1952.

24 "Survey of Advantages and Problems Associated with Transpiration Cooling and Film Cooling of Gas-Turbine Blades," by E. R. G. Eckert and J. B. Esgar, NACA RM E50K15, 1951.

25 "One-Dimensional Calculation of Flow in a Rotating Passage with Ejection Through a Porous Wall," by E. R. G. Eckert, J. N. B. Livingood, and E. I. Prasse, NACA TN 3408, 1955.

26 "Some NACA Investigations of Heat-Transfer Characteristics of Cooled Gas-Turbine Blades," by H. H. Ellerbrock, Jr., paper presented at General Discussion on Heat Transfer, The Institution of Mechanical Engineers, (London) and ASME (New York), Conference (London), September 11–13, 1951.

27 "Analysis of Spanwise Temperature Distribution in Three Types of Air-Cooled Turbine Blades," by J. N. B. Livingood and W. B. Brown, NACA Report 994, 1950.

28 "Procedure for Calculating Turbine Blade Temperatures and Comparison of Calculated With Observed Values for Two Stationary Air-Cooled Blades," by W. B. Brown, H. O. Slone, and H. T. Richards, NACA RM E52H07, 1952.

29 "A Summary of Basic Heat Transfer and Flow Friction Design Data for Plain Plate-Fin Heat Exchanger Surfaces," by W. M. Kays and S. H. Clark, Technical Report No. 17, prepared under Contract N6-ONR-251 Task Order 6 (NR-090-104) for Office of Naval Research, Department of Mechanical Engineering, Stanford University, Stanford, Calif., August 15, 1953.

30 "A Simplified Theory of Porous Wall Cooling," by W. D. Rannie, Progress Report No. 4-50, Power Plant Laboratory Project No. MX801, Jet Propulsion Laboratory C.I.T., November 24, 1947 (AMC Contract No. W-535-ac-20260, Ordnance Department Contract No. W-04-200-ORD-455).

31 "A Theoretical and Experimental Investigation of Rocket-Motor Sweat Cooling," by Joseph Friedman, *Journal of the American Rocket Society*, No. 79, December, 1949, pp. 147–154.

32 "Investigations of Air-Cooled Turbine Rotors for Turbojet Engines. II—Mechanical Design, Stress Analysis, and Burst Test of Modified J33 Split-Disk Rotor," by R. H. Kemp and M. L. Moseson, NACA RM E51J03, 1952.

33 "Investigations of Air-Cooled Turbine Rotors for Turbojet Engines. I—Experimental Disk Temperature Distribution in Modified J33 Split-Disk Rotor at Speeds up to 6000 Rpm," by W. B. Schramm and R. R. Ziemer, NACA RM E51I11, 1952.

34 "Investigations of Air-Cooled Turbine Rotors for Turbojet Engines. III—Experimental Cooling-Air Impeller Performance and Turbine Rotor Temperatures in Modified J33 Split-Disk Rotor up to Speeds of 10,000 Rpm," by A. J. Nachtigall, C. F. Zalabak, and R. R. Ziemer, NACA RM E52C12, 1952.

35 "Heat-Transfer and Operating Characteristics of Aluminum Forced-Convection and Stainless-Steel Natural Convection Water-Cooled Single-Stage Turbines," by J. C. Freche and A. J. Diaguila, NACA RM E50D03a, 1950.

36 "The Possibilities of the Gas Turbine for Aircraft Engines," by E. Schmidt, Reports and Transactions No. 489, GDC 2504T, British M.O.S.

37 "Application of Internal Liquid Cooling to Gas-Turbine Rotors," by S. Alpert, R. E. Grey, and D. D. Drake, Trans. ASME, vol. 78, 1956, pp. 1257–1266.

38 "Experiments on Mixed-, Free- and Forced-Convective Heat Transfer Connected with Turbulent Flow Through a Short Tube," by E. R. G. Eckert, A. J. Diaguila, and A. N. Curren, NACA TN 2974, 1953.

39 "Convective Heat Transfer for Mixed, Free, and Forced Flow Through Tubes," by E. R. G. Eckert and A. J. Diaguila, Trans. ASME, vol. 76, 1954, pp. 497–504.

40 "Free-Convection Effects on Heat Transfer for Turbulent Flow Through a Vertical Tube," by E. R. G. Eckert, A. J. Diaguila, and J. N. B. Livingood, NACA TN 3584, 1955.

41 "Review of Experimental Investigations of Liquid-Metal Heat Transfer," by Bernard Lubarsky and S. J. Kaufman, NACA TN 3336, 1955.

42 "Heat-Transfer Problems of Liquid-Cooled Gas-Turbine Blades," by Henry Cohen and F. J. Bayley, paper for presentation to The Institution of Mechanical Engineers, London, England, 1955.

43 "Free Convection in an Open Thermosiphon, with Special Reference to Turbulent Flow," by B. W. Martin, Proceedings of the Royal Society of London, series A, vol. 230, 1955, pp. 502–530.

44 "A Novel Cooling Method for Gas Turbine," by Edward Burke and G. A. Kemeny, Trans. ASME, vol. 77, 1955, pp. 187–195.

45 "Turbojet Afterburning Without Any Afterburner," by H. E. Schmitt, *Aeronautical Engineering Review*, vol. 9, December, 1950, pp. 18–24.

Generalized Optimal Heat-Exchanger Design[1]

By D. H. FAX[2] and R. R. MILLS, JR.[3]

There are many different ways of formulating the problem of the design of an optimum heat exchanger (or of any plant element or ensemble of elements) depending on what measure of performance is to be extremized and under what constraints the optimum is to be effected. A large number of such solutions exist in the literature but they often appear to have little in common, each variation in constraints apparently generating a brand new problem. Based on the use of Lagrangian multipliers, a generalized method is here developed which shows perhaps more clearly than before the interrelationship between different problems of this class. The method is particularly useful in those problems where several variables are to be optimized simultaneously. In the paper, the method is exemplified by the solution of three problems in gas-turbine regenerator design, at least two of which are believed to be new.

Nomenclature

The following nomenclature is used in the paper:

A = heat-transfer area, sq ft
c = ratio of air-side free-flow area to air-side frontal area, see Fig. 1
c_p = constant pressure specific heat, Btu/lb F
c_1, c_2, c_3, c_4, c_5 = functions of χ, τ, η_T, and η_c, defined in Equations [25] through [29]
d = hydraulic diameter of flow passages
E_t = cycle thermal efficiency
f = fanning friction factor
F = ratio of heat-exchanger mean temperature difference to that which would obtain in counterflow
g_c = $32.174 \times (3600)^2$ lb ft/hr^2 lb$_f$
G = mass velocity, W/S, lb/hr sq ft
h = heat-transfer coefficient, Btu/hr sq ft F
$j = \left[\frac{h}{c_p G}\right] \mathrm{Pr}^{2/3}$
$J\left[\frac{\ ,\ ,\ }{\ ,\ ,\ }\right] = \frac{\partial(\ ,\ ,\ )}{\partial(\ ,\ ,\ )}$ = Jacobian operator
k = isentropic exponent
M = Mach number based on frontal areas, $W_a v_a / XZ\sqrt{g_c k p_2 v_a}$ and $W_g v_g / YZ\sqrt{g_c k p_1 v_g}$ for air and gas sides, respectively
NTU = hA/Wc_p, number of transfer units, with subscript denoting air or gas side
p = pressure, psf, with subscripts as denoted in Fig. 2
$(\Delta p/p)_{a,\,g,\,c}$ = fractional pressure drops on air and gas sides of regenerator and in combustor
$(\Delta p/p)_T$ = total fractional pressure drop, defined in Equation [11]
Pr = Prandtl number
S = free-flow area, sq ft
T = temperatures, deg R, with subscripts as denoted in Fig. 2
v = specific volume, cu ft/lb
W = mass flow rate, lb/hr
x = $1 - \epsilon$
y = $\ln x$
X, Y, Z = regenerator dimensions, ft, as defined in Fig. 1
Γ = a dimensionless grouping defined in Equation [34]
ϵ = regenerator effectiveness
η = $2j/f$
η_c, η_T = adiabatic efficiencies of compressor and turbine
λ = a Lagrangian multiplier
τ = T_4/T_1, ratio of turbine-inlet to compressor-inlet temperatures
χ = $(p_2/p_1)^{\frac{k-1}{k}}$, compressor isentropic temperature ratio

Introduction

A number of writers have treated the problem of the design of optimal heat exchangers, a few of which are given in references (1–4).[4] The problems differ according to what measure of performance is to be extremized and under what types of constraints the optimum is to be effected. A study of the literature may give the impression that these problems have very little in common and that each additional variation in the constraints must perforce generate an entirely new problem.

Lagrange's method of multipliers is used in many branches of applied mathematics and physics to deal with the problem of extremizing a function of a number of variables which are, in turn, related by a number of equations of constraint. It is the purpose of this paper to show the application of this method to the optimal design of gas-turbine regenerators, and in particular to show the advantages provided by the method in revealing the points of similarity in a class of related problems.

The Method of Lagrangian Multipliers

Although a derivation of this method can be found in a number of places (e.g., 5, 6, 7), the form of the result to be used here is sufficiently different to warrant a special derivation.

Given a function of n variables

$$f = f(x_1, x_2, \ldots, x_n) \quad [1]$$

not all of which are independent but are related by m equations of constraint

$$\psi_j(x_1, x_2, \ldots, x_n) = 0 \quad [2]$$

$j = 1, 2, \ldots, m$; $m < n$. It is presumed that f and all the ψ's are differentiable throughout the range of interest.

The "conventional" method of extremizing the function f, presuming an extremum to exist, is to combine the ψ's with f so as to

[1] Based on an essay submitted to the Faculty of Engineering of The Johns Hopkins University, Baltimore, Md., in partial fulfillment of the requirements for the degree MSc Eng., May, 1954.

[2] Advisory Engineer, Commercial Atomic Power Activity, Westinghouse Electric Corporation, Pittsburgh, Pa.; formerly, Assistant Professor of Mechanical Engineering, The Johns Hopkins University, Baltimore, Md. Assoc. Mem. ASME.

[3] Research Assistant in Mechanical Engineering, The Johns Hopkins University, Baltimore, Md. Assoc. Mem. ASME.

Contributed by the Heat Transfer Division and presented at a joint session with the Gas Turbine Power Division at the Semi-Annual Meeting, Cleveland, Ohio, June 17–21, 1956, of The American Society of Mechanical Engineers.

Note: Statements and opinions advanced in papers are to be understood as individual expressions of their authors and not those of the Society. Manuscript received at ASME Headquarters, November 16, 1955. Paper No. 56—SA-19.

[4] Numbers in parentheses refer to the Bibliography at the end of the paper.

eliminate m of the x's, leaving f as a function of $n - m$ independent variables, then to equate the partial derivative of f with respect to each of the remaining x's to zero, and finally to solve this set of simultaneous equations for the optimal x's. This method can lead to some awkward algebra, particularly when the ψ's involve transcendental functions; more important, it can obscure the relationships among a class of problems which differ in only one (or a few) of the equations of constraint. The method to be presented frequently can obviate, or at least reduce, both of these difficulties.

A necessary condition for an extremum in f is that, at the extremum

$$df = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i} dx_i = 0 \quad [3]$$

Since the ψ's are equal to zero

$$d\psi_j = \sum_{i=1}^{n} \frac{\partial \psi_j}{\partial x_i} dx_i = 0 \quad [4]$$

Multiplying each $d\psi_j$ by λ_j, the Lagrangian multipliers, and adding all these products to Equation [3]

$$\sum_{i=1}^{n} \frac{\partial f}{\partial x_i} dx_i + \sum_{j=1}^{m} \lambda_j \sum_{i=1}^{n} \frac{\partial \psi_j}{\partial x_i} dx_i = 0$$

which, upon rearrangement, reads

$$\sum_{i=1}^{n} \left[\frac{\partial f}{\partial x_i} + \sum_{j=1}^{m} \lambda_j \frac{\partial \psi_j}{\partial x_i} \right] dx_i = 0 \quad [5]$$

Of the n brackets in Equation [5], m can be made zero by suitable choice of the m λ's; the remaining $n - m$ brackets also must be separately zero since $n - m$ of the x's are independent. Consequently we have n equations of the type

$$\frac{\partial f}{\partial x_i} + \sum_{j=1}^{m} \lambda_j \frac{\partial \psi_j}{\partial x_i} = 0; \quad i = 1, 2, \ldots, n \quad [6]$$

which, together with the m Equations of Constraint [2], can be solved explicitly for the n optimal x's and the m λ's.

The elimination of the m λ's can be performed at once. Choosing any m of Equations [6], say, the first m, and solving in the usual manner, we obtain

$$\lambda_j = - \frac{J\begin{bmatrix} \psi_1, \ldots, \psi_{j-1}, f, \psi_{j+1}, \ldots, \psi_m \\ x_1, x_2, \ldots, x_m \end{bmatrix}}{J\begin{bmatrix} \psi_1, \ldots, \psi_m \\ x_1, \ldots, x_m \end{bmatrix}} \quad [7]$$

where J denotes the Jacobian determinantal operator. It is to be noted that in order to effect these solutions, the denominator of Equation [7], common to all m of the λ's, must not be identically zero. This condition, that the mth-order determinant

$$J\begin{bmatrix} \psi_1, \ldots, \psi_m \\ x_1, \ldots, x_m \end{bmatrix} \not\equiv 0 \quad [8]$$

limits the choice as to which set of m of the n Equations [6] can be used to solve for the λ's. Now it can be shown that Equation [8] is both a necessary and sufficient condition that the equations of constraint be functionally independent with respect to the variables $x_1, \ldots, x_m$. Hence, if no set of m of the x's can be chosen so as to satisfy the Inequality [8], then at least one of the equations of constraint can be formed by some combination of the others. If this be so, then the problem stated by Equations [1] and [2] must be restated in terms of (at least) one less ψ and (at least) one more degree of freedom.

In order to exclude the trivial case $\lambda_j \equiv 0$, the numerators of Equations [7] cannot all be zero. Thus it is further necessary that the m x's be chosen such that not all of the $\partial f/\partial x_j$ are identically zero.

Having satisfied these two conditions, the solutions for the λ's may be substituted into the remaining $n - m$ of Equations [6], a typical one of which will look like

$$J\begin{bmatrix} \psi_1, \ldots, \psi_m \\ x_1, \ldots, x_m \end{bmatrix} \frac{\partial f}{\partial x_r} - \sum_{j=1}^{m} J\begin{bmatrix} \psi_1, \ldots, \psi_{j-1}, f, \psi_{j+1}, \ldots, \psi_m \\ x_1, \ldots, x_m \end{bmatrix} \frac{\partial \psi_j}{\partial x_r} = 0 \quad [9]$$

where $r = m + 1, m + 2, \ldots, n$. This form is recognized as the expansion by cofactors of the $(m + 1)$th-order determinant

$$J\begin{bmatrix} \psi_1, \ldots, \psi_m, f \\ x_1, \ldots, x_m, x_r \end{bmatrix} = 0 \quad [10]$$

The solution for the n optimal x's is given by the m equations of constraint and the $n - m$ determinantal Equations [10]. The λ's do not appear explicitly in the final formulation of the method. The algebra required to solve the n simultaneous equations may still be quite tedious. Nevertheless, the present method should be simpler than the conventional approach in many cases because the final solution can always be effected, if necessary, by numerical or graphical means and no calculus operations remain to be performed on the solutions so found.

It is obvious that if one or more of the equations of constraint can be written in the form

$$\psi_j(x_1, x_2, \ldots, x_n) = \text{const}_j$$

the preceding work remains unchanged. It is further obvious that the interchange of any columns within any of the determinants denoted by Equations [10] does not change the value of the determinant. Thus it is clear that the role of the function of f can be interchanged with that of any one of the constraints which can be written as above without affecting the final result, providing, of course, that a set of m x's satisfying the Inequality [8] can be found in each case. In other words, given a set of $m + 1$ independent measures of performance which are functions of a larger set of design variables which may be optimized, then regardless of which one of the performance criteria is to be extremized (the other m being fixed), the optimal relations among the design variables is the same. This important result, which will be exemplified in the following, is not so apparent from the conventional approach.

Applications of the Method

The Lagrangian method will be exemplified by the solution of three problems relating to the design of gas-turbine heat exchangers. In each case, the exchanger is presumed to be of the cross-flow plate-fin type, Fig. 1, and the thermodynamic process of which it is a part is shown in Fig. 2. The first problem, solved in somewhat different form by Aronson (2), is to find the optimal division between free-flow areas and NTU on the two sides of the exchanger for a prescribed over-all effectiveness and cycle pressure ratio. The second problem will lift the prescription of the over-all effectiveness, and the third problem will lift the prescription of the cycle pressure ratio as well.

Optimal Exchanger Proportions for Prescribed Effectiveness and Cycle Pressure Ratio

It is well known (2) that the irreversibilities caused by pressure

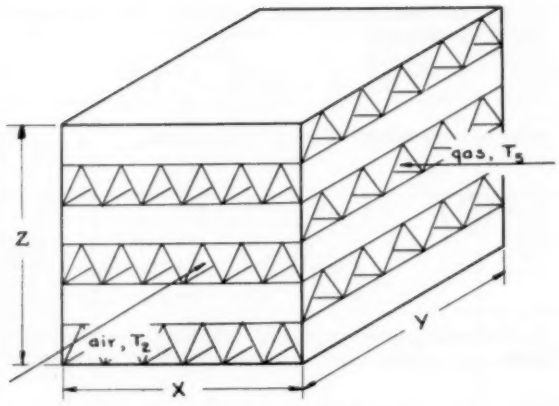


FIG. 1 CROSS-FLOW PLATE-FIN REGENERATOR

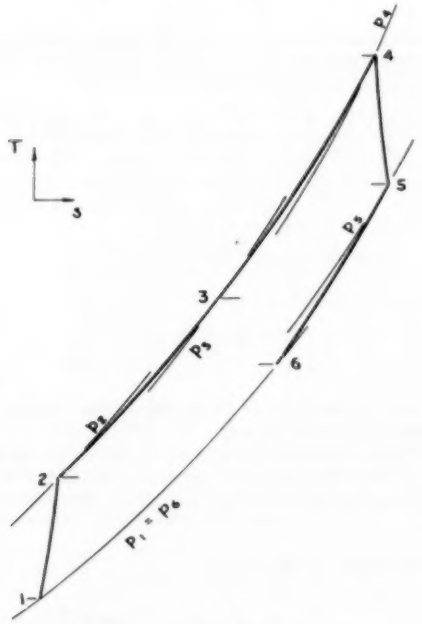


FIG. 2 REGENERATIVE GAS-TURBINE CYCLE
(Path 1–2, compressor process; 2–3, regenerator air side; 3–4, combustor; 4–5, turbine; 5–6, regenerator gas side.)

drops in the gas-turbine process in Fig. 2 can be expressed by the sum

$$\left(\frac{\Delta p}{p}\right)_T = \left(\frac{\Delta p}{p}\right)_a + \left(\frac{\Delta p}{p}\right)_g + \left(\frac{\Delta p}{p}\right)_c \quad \ldots\ldots [11]$$

where the term on the left is proportional to the total irreversibility occasioned by the pressure drops and the terms on the right are the fractional drops in the air side and gas side of the exchanger and in the combustor, respectively. For the air side of the regenerator, one may write

$$\left(\frac{\Delta p}{p}\right)_a = \frac{\mathrm{Pr}^{2/3} W^2 v_a (\mathrm{NTU})_a}{\eta g_c p_2 S_a^2} \quad \ldots\ldots [12]$$

where v_a is the mean specific volume on the air side, $S_a = c\,XZ$ is the free-flow area for the air, and $\eta = (2j/f)$ relates the heat-transfer and friction characteristics of the exchanger surface. Following Aronson (2), it will be assumed that variations in η (which depend on Reynolds number) may be ignored as a second-order effect. It will be further assumed, but only for convenience in symbolism, that η is the same on both sides of the exchanger and it has been tacitly assumed in Equation [11] that the mass rate of flow W is the same for both air and gas.

Using the p-v-T relation for perfect gases, Equation [12] may be rewritten in the form

$$\left(\frac{\Delta p}{p}\right)_a = \left(\frac{\mathrm{Pr}^{2/3} k}{\eta}\right) \frac{\mathbf{M}_a^2\,\mathrm{NTU}_a}{c^2} \quad \ldots\ldots [13]$$

where k is the isentropic exponent and $\mathbf{M}_a$ is the Mach number based on the total frontal area XZ and the mean temperature on the air side. Specification of $\mathbf{M}_a$ and the analogously defined $\mathbf{M}_g$ specifies in a convenient manner the two frontal areas. These Mach numbers will be quite low, in the neighborhood of 0.01 to 0.05, for most plant types. The choice of specific values would depend upon cost factors beyond the scope of the present discussion.

For the present problem, therefore, we wish to minimize the function

$$\left(\frac{\Delta p}{p}\right)_T = \left(\frac{\mathrm{Pr}^{2/3} k}{\eta}\right)\left[\frac{\mathbf{M}_a^2\,\mathrm{NTU}_a}{c^2} + \frac{\mathbf{M}_g^2\,\mathrm{NTU}_g}{(1-c)^2}\right] + \left(\frac{\Delta p}{p}\right)_c \quad \ldots\ldots [14]$$

subject to the constraint that a specified amount of heat shall be transferred. For the case of equal heat-capacity rates on the two sides, this constraint can be expressed as

$$\psi_1 = \frac{1}{\mathrm{NTU}_a} + \frac{1}{\mathrm{NTU}_g} = \left(\frac{1-\epsilon}{\epsilon}\right) F \quad \ldots\ldots [15]$$

where F, the ratio of the actual mean temperature difference to that which would obtain in counterflow, is itself a function of the effectiveness ϵ and the specific flow arrangement (8). The function ψ_1 is thus a measure of the over-all thermal resistance of the exchanger.

TABLE 1 MATRIX FOR THE OPTIMIZATION OF REGENERATOR PROPORTIONS

	$\frac{\partial}{\partial (\mathrm{NTU})_a}$	$\frac{\partial}{\partial (\mathrm{NTU})_g}$	$\frac{\partial}{\partial c}$
$\left(\frac{\Delta p}{p}\right)_T$	$\frac{\mathbf{M}_a^2}{c^2}$	$\frac{\mathbf{M}_g^2}{(1-c)^2}$	$\frac{2\,\mathbf{M}_a^2\,\mathrm{NTU}_a}{c^3} - \frac{2\,\mathbf{M}_g^2\,\mathrm{NTU}_g}{(1-c)^3}$
ψ_1	$-\frac{1}{\mathrm{NTU}_a^2}$	$-\frac{1}{\mathrm{NTU}_g^2}$	0

The partial derivatives to be used in the solution of the problem are arranged in matrix form in Table 1. The determinants which are to be set equal to zero will be formed from the rows and columns of this matrix. It will be seen that either NTU_a or NTU_g may be chosen as the variable satisfying Inequality [8]. The simultaneous solution of the equations

$$J\left[\frac{(\Delta p/p)_T,\ \psi_1}{\mathrm{NTU}_a,\ \mathrm{NTU}_g}\right] = 0 \quad \text{and} \quad J\left[\frac{(\Delta p/p)_T,\ \psi_1}{\mathrm{NTU}_a,\ c}\right] = 0$$

yields the result that at the optimum

$$\frac{\mathrm{NTU}_a}{\mathrm{NTU}_g} = \sqrt{\frac{\mathbf{M}_g}{\mathbf{M}_a}} \quad [16]$$

and

$$c = 1/\left(1 + \sqrt{\mathbf{M}_g/\mathbf{M}_a}\right) \quad [17]$$

Combining these results with Equations [14] and [15], we further find that at the optimum design

$$\mathrm{NTU}_a = \frac{\epsilon}{(1-\epsilon)F}\left[1 + \sqrt{\frac{\mathbf{M}_g}{\mathbf{M}_a}}\right] \quad [18]$$

$$\mathrm{NTU}_g = \frac{\epsilon}{(1-\epsilon)F}\left[1 + \sqrt{\frac{\mathbf{M}_a}{\mathbf{M}_g}}\right] \quad [19]$$

and

$$\left(\frac{\Delta p}{p}\right)_T = \left(\frac{\Delta p}{p}\right)_c + \frac{k\mathrm{Pr}^{2/3}}{\eta}\frac{\epsilon}{(1-\epsilon)F}\left[\sqrt{\mathbf{M}_a} + \sqrt{\mathbf{M}_g}\right]^4 \quad [20]$$

As an example of the interchangeability of the constraint with the function to be extremized, it is to be noted that the identical results are obtained whether the problem be to minimize the over-all $\Delta p/p$ for a prescribed over-all thermal resistance, or to minimize the thermal resistance for a specified over-all $\Delta p/p$.

The specification of the two frontal area Mach numbers, together with the optimal Relations [17], [18], and [19], does not completely fix the design of the two sides of the exchanger. Given the heat-transfer and friction characteristics of the surfaces as functions of Reynolds number, we are yet free to choose any one of the following dimensions: X, Y, Z, d_a, or d_g. It is easily shown that if one of the hydraulic diameters be chosen for specification, then the over-all volume decreases continuously as that diameter is made smaller while the two frontal areas, which are kept constant, become more and more elongated in the Z (nonflow) direction. This can, of course, lead to awkward ducting even though the volume of the exchanger itself can become relatively small. The completion of this problem, consequently, depends on an interrelation of cost factors and plant layout problems which, in turn, depend on the nature of the specific application. This has previously been pointed out in another context by Aronson (9).

Optimal Effectiveness for Prescribed Pressure Ratio

In this second problem, we shall determine the regenerator effectiveness (as well as its proportions) which maximizes the gas-turbine-cycle thermal efficiency for prescribed values of the over-all pressure ratio and over-all temperature ratio. Our first task will be to write the thermal efficiency in a form amenable to the analysis which follows.

Under the assumption of constant specific heat, the cycle thermal efficiency (by reference to Fig. 2) is given by

$$E_t = \frac{(T_4 - T_5) - (T_2 - T_1)}{(T_4 - T_2)}$$

which, upon combination with the regenerator effectiveness

$$\epsilon = (T_3 - T_2)/(T_5 - T_2)$$

is written as

$$E_t = \frac{(T_4 - T_5) - (T_2 - T_1)}{\epsilon(T_4 - T_5) + (1-\epsilon)[(T_4 - T_1) - (T_2 - T_1)]} \quad [21]$$

The temperature rise in the compressor is

$$T_2 - T_1 = T_1(\chi - 1)/\eta_c \quad [22]$$

where

$$\chi = (p_2/p_1)^{\frac{k-1}{k}}$$

and η_c is the compressor adiabatic efficiency. The temperature drop in the turbine is

$$T_4 - T_5 = \eta_T T_4\left(1 - \frac{1}{\chi_T}\right)$$

where η_T is the turbine adiabatic efficiency and

$$\frac{1}{\chi_T} = \left(\frac{p_5}{p_4}\right)^{\frac{k-1}{k}} = \left(\frac{p_1 + \Delta p_g}{p_2 - \Delta p_a - \Delta p_c}\right)^{\frac{k-1}{k}} = \left(\frac{p_1}{p_2}\right)^{\frac{k-1}{k}}\left[\frac{1 + (\Delta p/p)_g}{1 - (\Delta p/p)_a - (\Delta p/p)_c}\right]^{\frac{k-1}{k}}$$

The fractional pressure drops given in the last expression are quite small compared with unity, so to within the first order in $(\Delta p/p)$

$$T_4 - T_5 = \eta_T T_4\left[1 - \frac{1 + \left(\frac{k-1}{k}\right)\left(\frac{\Delta p}{p}\right)_T}{\chi}\right] \quad [23]$$

where $(\Delta p/p)_T$ is the sum of the three fractional pressure drops.

Inserting Equations [22] and [23] in [21] and again neglecting terms of order $(\Delta p/p)^2$ and higher, we arrive at

$$\frac{1}{E_t} = \frac{1}{c_1}\left[c_3 - c_4\epsilon + \frac{c_2}{c_1}(c_3 - c_4\epsilon)\left(\frac{\Delta p}{p}\right)_T\right] \quad [24]$$

where

$$c_1 = (\chi - 1)(\tau\eta_T\eta_c - \chi) \quad [25]$$

$$c_2 = \tau\eta_T\eta_c(k-1)/k \quad [26]$$

$$c_3 = \chi[\eta_c(\tau - 1) - (\chi - 1)] \quad [27]$$

$$c_4 = c_2 - \tau\eta_T\eta_c(\chi - 1) \quad [28]$$

$$c_5 = c_1 + c_4 \quad [29]$$

and $\tau = T_4/T_1$, the over-all temperature ratio. Equation [24] is the function to be minimized.

The relation between ϵ and $(\Delta p/p)_T$ is given by the equations of constraint. The function which was minimized in the preceding problem becomes one of the constraints in the present problem, namely

$$\psi_1 = \frac{\mathbf{M}_a^2\,\mathrm{NTU}_a}{c^2} + \frac{\mathbf{M}_g^2\,\mathrm{NTU}_g}{(1-c)^2} + \frac{\eta}{k\mathrm{Pr}^{2/3}}\left[\left(\frac{\Delta p}{p}\right)_c - \left(\frac{\Delta p}{p}\right)_T\right] = 0 \quad [30]$$

while

$$\psi_2 = \frac{1}{\mathrm{NTU}_a} + \frac{1}{\mathrm{NTU}_g} - \left(\frac{1-\epsilon}{\epsilon}\right)F = 0 \quad [31]$$

Since F is a function of ϵ, which is now no longer fixed, we must include a relation between them. This relation depends on the specific flow arrangement and upon the degree of mixing in the directions normal to their flow for each of the two fluids. For the purpose of the example we choose a single-pass cross-flow exchanger in which one of the fluids is assumed to be mixed, the other unmixed. From Equation [11] of reference (8), the relation for this case can be expressed (when the two heat-capacity rates are equal) by

$$\psi_3 = F + \frac{\epsilon}{(1-\epsilon)\ln\left[1 + \ln(1-\epsilon)\right]} = 0 \quad [32]$$

TABLE 2 MATRIX FOR THE OPTIMIZATION OF EFFECTIVENESS AND CYCLE PRESSURE RATIO

	$\frac{\partial}{\partial F}$	$\frac{\partial}{\partial(\Delta p/p)_T}$	$\frac{\partial}{\partial(\mathrm{NTU})_a}$	$\frac{\partial}{\partial(\mathrm{NTU})_g}$	$\frac{\partial}{\partial c}$	$\frac{\partial}{\partial \epsilon}$	$\frac{\partial}{\partial \chi}$
$\frac{1}{E_t}$	0	$\frac{c_3}{c_1}(c_3 - c_5\epsilon)$	0	0	0	$-\frac{c_4}{c_1} - \frac{c_2 c_3}{c_1^2}\left(\frac{\Delta p}{p}\right)_T$	$\frac{\partial(1/E_t)}{\partial \chi}$
ψ_1	0	$-\frac{\eta}{k\,Pr^{2/3}}$	$\frac{M_a^2}{c^2}$	$\frac{M_g^2}{(1-c)^2}$	$\frac{2\,M_a^2\,NTU_a}{c^3} - \frac{2\,M_g^2\,NTU_g}{(1-c)^3}$	0	0
ψ_2	$-\frac{1-\epsilon}{\epsilon}$	0	$-\frac{1}{NTU_a^2}$	$-\frac{1}{NTU_g^2}$	0	$\frac{F}{\epsilon^2}$	0
ψ_3	1	0	0	0	0	$\frac{\ln(1+\ln x) + \frac{1-x}{1+\ln x}}{[x\ln(1+\ln x)]^2}$	0

NOTE: The first six columns apply to the second problem in the text; the entire matrix applies to the third problem.

Equations [24], [30], [31], and [32] complete the statement of this second problem. Table 2 shows the matrix of partial derivatives appropriate to the method used herein. It is evident that one set of variables satisfying Inequality [8] is $F, (\Delta p/p)_T, \mathrm{NTU}_a$.

Simultaneous solution of the equations

$$J\left[\frac{1/E_t, \psi_1, \psi_2, \psi_3}{F, (\Delta p/p)_T, \mathrm{NTU}_a, \mathrm{NTU}_g}\right] = 0$$

and

$$J\left[\frac{1/E_t, \psi_1, \psi_2, \psi_3}{F, (\Delta p/p)_T, \mathrm{NTU}_a, c}\right] = 0$$

yields results duplicating those of the preceding problems as given by Equations [16] through [20]. This indicates that, under the assumptions made here, the optimal proportions of the regenerator depend on ϵ in a manner that is independent of whether ϵ has itself been optimized or not. The optimal ϵ is given by the solution of the fourth determinantal equation

$$J\left[\frac{1/E_t, \psi_1, \psi_2, \psi_3}{F, (\Delta p/p)_T, \mathrm{NTU}_a, \epsilon}\right] = 0$$

which can be written as

$$y = -1 + \left[\frac{1 + (1+y)\ln(1+y) + \left(\frac{c_2}{c_3} - 1\right)e^{-y}}{\frac{c_1 c_4}{c_2 c_3} + \left(\frac{\Delta p}{p}\right)_a}\right]\Gamma \quad \ldots\ldots [33]$$

where

$$y = \ln x = \ln(1-\epsilon)$$

and

$$\Gamma = \frac{k\mathrm{Pr}^{2/3}}{\eta}\left[\sqrt{M_a} + \sqrt{M_g}\right]^4 \ldots\ldots\ldots [34]$$

Equation [33] cannot be solved explicitly for y (or ϵ) but the form given here is particularly suitable for iterative solution. An assumed value of y inserted on the right yields an improved value on the left and the process is repeated to convergence. From Equation [32] it can be shown that as F varies from zero to unity, ϵ covers the range $(e-1)/e$ to zero; hence y has the limits -1 and zero. The iteration indicated by Equation [33] converges rapidly regardless of which end of the range is chosen for the first step in the process.

Having found the optimal value of ϵ from Equation [33], the optimal $(\Delta p/p)_T$ can be evaluated by combining Equations [20] and [32]; it is

$$(\Delta p/p)_T = (\Delta p/p)_a - \Gamma \ln[1 + \ln(1-\epsilon)] \ldots\ldots [35]$$

The fractional pressure drop in the combustor has of course been assumed to be a prescribed constant.

Optimal Pressure Ratio

If we now set the cycle pressure ratio free and wish to find its optimal value (for prescribed over-all temperature ratio τ), we add a fourth degree of freedom to the system described in the previous problem. Examination of Table 2 shows that the preceding work remains applicable to the present case; the optimal proportions of the regenerator are still given by Equations [16] through [20] and the optimal effectiveness by the soluton of Equation [33]. Hence the present problem requires the solution of only one new equation

$$J\left[\frac{1/E_t, \psi_1, \psi_2, \psi_3}{F, (\Delta p/p)_T, \mathrm{NTU}_a, \chi}\right] = 0$$

Because of the three zero elements in the last column of the matrix of Table 2, this equation reduces to

$$J\left[\frac{\psi_1, \psi_2, \psi_3}{F, (\Delta p/p)_T, \mathrm{NTU}_a}\right]\frac{\partial(1/E_t)}{\partial \chi} = 0$$

The first term of this product has already been shown to be nonzero, hence this third problem is solved by setting

$$\frac{\partial(1/E_t)}{\partial \chi} = 0$$

Performing the indicated differentiation on Equation [24] and using Equations [25] through [29], we arrive at an equation of the fourth degree in the optimal value of χ. The equation can be solved, however, by a rapidly converging iteration in the following form

$$[\tau\eta_T(2\epsilon - 1) + (\tau - 1)(1 - \epsilon)]\chi^2 - 2\tau\eta_T(2\epsilon - 1)\chi + \tau\eta_T[(2\epsilon - 1) - \eta_c(\tau - 1)(1 - \epsilon)] = \left(\frac{k-1}{k}\right)\left(\frac{\Delta p}{p}\right)_T \tau\eta_T \left\{(2\epsilon - 1)\left[\frac{\tau\eta_T\eta_c - \chi(2\chi - 1)}{\tau\eta_T\eta_c - \chi}\right] + \frac{(1-\epsilon)(\tau-1)}{\chi - 1}\eta_c\left[\frac{\tau\eta_T\eta_c - \chi(2\chi-1)}{\tau\eta_T\eta_c - \chi} + \chi\right]\right\} \quad [36]$$

The procedure is as follows: A reasonable value of χ is assumed and the corresponding optimal ϵ and $(\Delta p/p)_T$ are computed by the equations given in the first two problems. With these substituted in Equation [36], together with the initial assumption of χ where it appears on the right of that equation, the resulting quadratic is readily solved for an improved value of χ and the process is repeated. Subsequent solutions of Equation [36] have been shown by trial to yield oscillating values of decreasing amplitude for χ; hence convergence can be hastened by judicious choices of the value of χ for the successive steps of the iteration.

Summary

The method of Lagrangian multipliers is presented in a form which does not require the explicit solution for the multipliers themselves. To optimize a set of n variables which are interrelated by m equations of constraint, it is only necessary instead to solve simultaneously a set of $n - m$ determinants of $(m + 1)$th order each set equal to zero, together with the m equations of constraint, in order to obtain directly the optimal values of the n variables.

Applied to problems of process and plant design, it is believed that this method can afford more insight in the synthesis of the mathematical statement of the problem and can be simpler in the execution of the solution than more widely used methods. In particular, it lends itself more readily to numerical evaluation of results where the equations of constraint are transcendental in form.

In the paper, the method is applied to the problem of simultaneously optimizing the cycle pressure ratio, and the regenerator effectiveness, pressure drops, and proportions for maximum thermal efficiency in a gas-turbine cycle of specified turbine and compressor-inlet temperatures. It becomes clear that the optimization of the regenerator proportions is independent of the optimization of the regenerator effectiveness and that both are independent of the optimization of the cycle pressure ratio.

While a particular function, Equation [32], was chosen to relate the effectiveness to the mean temperature difference, any other applicable relationship would work as well and such a change would involve only the last row of the matrix in Table 2. Similarly, if it were required that some function other than thermal efficiency be extremized (e.g., cost or weight), the change would involve only the first row of the matrix. It is believed that this "localization" of changes in the statement of such problems is a particular merit of the method described herein.

BIBLIOGRAPHY

1 "Heat Transmission," by W. H. McAdams, McGraw-Hill Book Company, Inc., New York, N. Y., third edition, 1954, chapter 15.

2 "Design of Regenerators for Gas-Turbine Service" by D. Aronson, Trans. ASME, vol. 72, 1950, pp. 967–978.

3 "Optimum Design of Gas Turbine Regenerators," by W. M. Rohsenow, T. R. Yoos, Jr., and J. F. Brady, ASME Paper No. 50—A-103.

4 "Review of Optimum Design of Gas-Turbine Regenerators," by D. Aronson, Trans. ASME, vol. 74, 1952, pp. 675–683.

5 "Differential and Integral Calculus," by R. Courant, Interscience Publishing Company, New York, N. Y., vol. 2, 1952.

6 "Advanced Calculus," by W. Kaplan, Addison-Wesley Press, Inc., Cambridge, Mass., 1952, pp. 128–136.

7 "Heat and Thermodynamics," by M. W. Zemansky, McGraw-Hill Book Company, Inc., New York, N. Y., third edition, 1951, pp. 428–429.

8 "Mean Temperature Difference in Design," by R. A. Bowman, A. C. Mueller, and W. M. Nagle, Trans. ASME, vol. 62, 1940, pp. 283–294.

9 Discussion by D. Aronson of "Heat-Transfer and Flow-Friction Characteristics of Some Compact Heat-Exchanger Surfaces; Part 2—Design Data for Thirteen Surfaces," by W. M. Kays and A. L. London, Trans. ASME, vol. 72, 1950, p. 1096.

Discussion

David Aronson.[5] The authors take a fresh approach to the problem of optimal heat-exchanger design. The writer appreciates the simplicity of treatment which the authors achieve by the use of Lagrangian multipliers and the solution of the equations in an ordered fashion by means of the Jacobian operator. The authors are to be commended for their clear presentation of the design parameters to be optimized:

1 Optimal exchanger proportions as regards balance between hot and cold-passage design.
2 Optimal effectiveness for prescribed pressure ratio.
3 Optimal pressure ratio.

The authors' results for the analysis of the first topic are in agreement with the ones the writer obtained (reference 2 of the paper) and he is grateful for this independent check of his previous work. The solution of the second topic appears to be correct as regards the formal mathematics, but seems in error in the selection of design relations. In the opinion of the writer, the third topic is not amenable to analytical treatment in the simple fashion chosen by the authors.

The writer recently submitted to a Review Committee of the ASME a possible solution to the second topic, in the nature of an approximation. He, therefore, was looking with some eagerness to the solution the authors propose, since they apparently endeavored to avoid approximations in their development. However, he was disappointed to find that they had glossed over the same troublesome spot as he had. The problem can be stated in these terms: The pressure ratio of a gas-turbine cycle has been set, say, by the turbine designer. Limitations also are imposed on the design of the heat exchanger. One can place a limit on the heat-exchanger volume, cross-sectional area for flow, or the total heat-transfer-surface area. The authors do not state clearly which limitation they have chosen. The treatment of the problem, however, indicates that they have specified a fixed cross-sectional area for flow. One must look at this restraint more closely. The correct statement of the condition should read: "Cross-sectional area for flow is fixed for a given net work." The authors set up the problem on the basis of: "Cross-sectional area for flow is fixed for a given working-fluid flow rate." Had the authors chosen to limit heat-transfer-surface area, the distinction between basing this limit on net work or on working-fluid flow rate would have been much less significant, because the optimal pressure drop would be about one fourth that for the limiting condition studied in the paper.

The authors' final equation is stated to be suitable for iterative solution; it cannot be solved explicitly. The iterative solution may or may not be possible for the general case of any type of heat-exchanger arrangement, counterflow, multipass crossflow, or single-pass crossflow. The condition chosen by the authors seems to the writer to be a strange one. They use the equation ([32] of the paper) for a single-pass, crossflow exchanger with fluid mixed on one side, unmixed on the other. For this case the maximum effectiveness, regardless of heat-exchanger size is only

[5] Consulting Engineer, Worthington Corporation, Harrison, N. J. Mem. ASME.

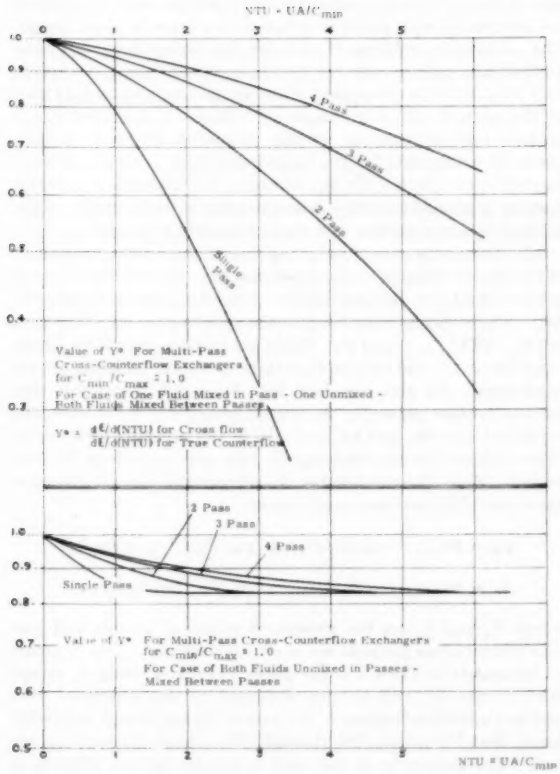


FIG. 3

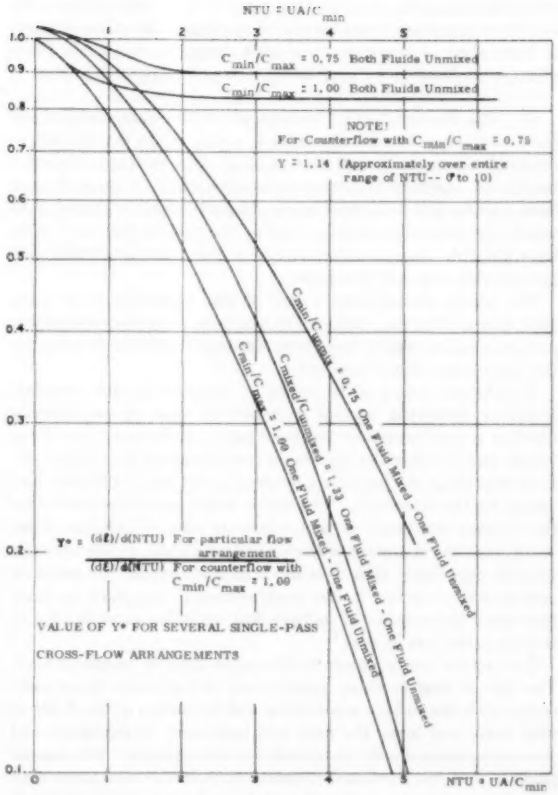


FIG. 4

about 63 per cent. This is such a low effectiveness that it is hard to consider it in a study of the optimum. The heat exchanger illustrated in Fig. 1 of the paper would seem to approach in its performance more closely to the case of fluids unmixed on both sides, for which case the effectiveness can approach 100 per cent as the surface area is increased without limit.

The writer arrived at the following solution of the problem as set forth by the authors

$$\Delta P/P_g + \Delta P/P_a = E_t \left(\frac{k}{k-1}\right)\left(\frac{T_b - T_2}{T_b}\right)\frac{(\text{NTU})\,(Y^*)}{(1+\text{NTU})^2} \quad \ldots\ldots\ldots [37]$$

where Y^* is a correction term to allow for the difference in slope of the ϵ versus NTU curve for the particular flow arrangement as compared with the counterflow arrangement for $C_{min}/C_{max} = 1.00$. For the crossflow arrangement used by the authors, Y would have a value of 0.51 at NTU = 2.0 and a value of 0.18 at NTU = 4.0.

The term NTU represents the over-all NTU of both sets of passages, as defined by

$$1/\text{NTU} = 1/\text{NTU}_a + 1/\text{NTU}_g$$

for the special case of $C_{min}/C_{max} = 1.00$.

This indicated optimum pressure-drop ratio is in error in the same fashion as would be the answer obtained by the authors in that no allowance is made for the reduction in net work due to the pressure loss in the heat exchanger.

For the case of heat-transfer-surface area being considered specified, the total pressure-drop ratio given by Equation [37] herewith, would be multiplied by a term r dependent on fin effectiveness and on the value of the exponents for the change of friction factor f and heat-transfer factor j with respect to Reynolds number. The term r has a value in the order of 0.25 for turbulent flow, with a slight increase or decrease in this value dependent on fin effectiveness and whether the flow is along smooth surfaces or across interrupted surfaces.

The third topic considered is even more vulnerable to the criticism that it does not hold to a fixed net work. Presumably one could modify the equations to adjust for the decrease in net work per unit of working-fluid flow. Such an analysis would have no practical interest since the change in plant cost with change in pressure ratio and, hence, of flow volumes would not be given consideration. The selection of optimum-pressure ratio is tied in with the design of compressor, turbine, combustion chamber, and ducts as well as with the design of the heat exchanger. Perhaps the techniques described by the authors could be applied to such an analysis. This would call for inclusion of capital costs, fuel costs, interest and amortization schedules, and so on. The distinction between this latter problem and the previous ones lies in the nature of the choices. Topics one and two can be treated purely on the basis of the thermodynamic consideration of the irreversibilities in the heat exchangers themselves. Topic three is primarily concerned with the optimum capital investment in the plant. The familiarity with the solutions indicated by proper treatment of topics one and two can aid materially in the solution of the problem posed by topic three. The authors have contributed constructively on the first two topics. It is felt that they

would enhance the value of their paper if they were to illustrate how their proposed solutions can be applied to the study implied in topic three, but not proceed with the solution. Topic three offers a challenge justifying another paper or several papers.

WALTER DASKIN.[6] The authors are to be congratulated for their lucid exposition of a technique which, while well known to mathematicians and engineers dealing with problems in solid mechanics, seems to have been virtually ignored by those of us in heat-transfer and fluid mechanics. In particular, the clarity with which the inter-relationships among the various extremal problems has been demonstrated is one of the most important and far-reaching points of this paper.

The writer having been aware of this particular work since 1954, has applied the technique to a number of cycle-optimization problems which would have been extremely tedious to solve by the more conventional methods.

A difficulty which often arises in connection with extremal problems involving several variables is that of establishing whether a true extremum (rather than a saddle-point) has been found, and whether the extremum is a maximum or a minimum. Equation [3] of the paper is a necessary, but not a sufficient condition, for the extremum. The writer would like to know whether the authors are aware of any relatively simple methods of investigating the character of the stationary point. It has been the writer's experience that it is often easiest actually to calculate values of the function in the neighborhood of the point in question when the number of independent variables exceeds two—a tedious procedure at best.

Two rather minor points in the paper deserve comment here. The first is that in many applications the pressure drops associated with the sudden contraction and expansion of the fluids as they enter and leave the heat exchanger may be significant and these quantities should be included in the analysis. The second point is that the fractional pressure drop in the combustor may often be expressed as a function of $T_4 - T_3$ and thus included in the analysis with little additional effort.

The paper clearly points to the need for systematic data on the performance of heat-exchanger surfaces. Despite the monumental effort by Kays and London[7] we are still in the state where we must often go through a great number of calculations, each assuming different types of surface, in order to find what particular surface would result in the best heat-exchanger design. While the present paper shows how to find the optimum configuration once a surface is chosen, it is still often necessary to perform the calculations for a number of surfaces in order to arrive at the best configuration. An example of this sort of calculation (in which the optimization technique which was used was the conventional one) is given.[8] It would be extremely pleasant for the designer to have some sort of analytic function describing the variation in heat-transfer and friction-drop performance when going from one kind of surface to another.

AUTHORS' CLOSURE

The authors thank Mr. Aronson and Mr. Daskin for their careful reviews of the paper and for the questions which they raise.

The prime purposes of the paper were to show that the method of Lagrangian multipliers can be simplified (by implicit elimination of the multipliers themselves) and to show that the method can be applied with profit to practical problems of plant design. The particular problems chosen for the exemplification of the method were only secondary; the choice was dictated to some extent by a desire to cover most of the interesting and useful facets of the method with a minimum of tedium. Consequently, the authors take no umbrage at Mr. Aronson's criticism of their choice of constraints for the second and third examples demonstrated in the paper. On the contrary, the question is welcome because it permits an added demonstration of the flexibility of the method to accommodate just such changes in constraints.

Mr. Aronson is correct in saying that in the original statement of the second problem we have assumed that the two frontal areas per unit fluid-flow rate are fixed by prescribing values for $\mathbf{M}_a$ and $\mathbf{M}_g$. The variables remaining in the problem are six: $(\Delta p/p)_T$, NTU_a, NTU_g, c, ϵ, and F. These are in turn related by Equations [30], [31], and [32], so that three of the six variables are independent. To accommodate Mr. Aronson's suggestion that he would rather prescribe the frontal areas per unit net work, one would set free $\mathbf{M}_a$ and $\mathbf{M}_g$ and add to the previous three equations (which remain unchanged) two new equations of constraint. Thus, the net number of independent variables remains the same. The two new equations are

$$\psi_4 = \mathbf{M}_a[c_1 - c_2(\Delta p/p)_T] = \chi\eta_c/F_a c_p T_1\sqrt{(g_c k p_2/v_a)}$$

$$\psi_5 = \mathbf{M}_g[c_1 - c_2(\Delta p/p)_T] = \chi\eta_c/F_g c_p T_1\sqrt{(g_c k p_1/v_g)}$$

where F_a and F_g are the prescribed values of air-side and gas-side frontal areas per unit net work, respectively. The elements of the matrix in Table 2 of the paper remain unchanged, except that the matrix itself becomes enlarged by two additional rows and two additional columns. As part of the solution, it is quickly found that Equations [16] through [20] remain the same as before. The remainder of the work is tedious but not difficult in essence.

Mr. Aronson is correct in saying that a crossflow exchanger in which one fluid is mixed and the other unmixed has a very low maximum effectiveness. Again, the choice was made on the basis of reasonable simplicity for the purpose of exposition. The exact relation between the functions F and ϵ in Equation [32] for the case of both fluids unmixed (8) is indeed a formidable one for our present purposes; nevertheless, it could very well be closely approximated, at least over an appropriate range of interest, by a less complicated analytical expression. It is worth repeating that such a change would affect only the third row of the matrix in Table 2 and none of the other work done to that point.

In answer to Mr. Aronson's last question, the method is equally applicable to the problem of minimizing cost, provided only that the problem can be stated in mathematical terms using differentiable functions. One of the authors has recently solved such a problem (in a field other than gas-turbine plants) in which there were ten variables related by five equations of constraint. The basic question is, of course, the proper balance between fixed and variable costs and, as expected, the answers are functions of plant capacity factor and rate of amortization as well as unit costs of equipment and fuel.

The authors are particularly pleased that Mr. Daskin has found this work useful. His question as to how to determine whether the stationary point is an extremum or a saddle point is well taken. Where f is a function of n x's, all of which are independent, the stationary point is a minimum if the $n \times n$ matrix $||\partial^2 f/\partial x_i \partial x_k||$ is positive definite; i.e., if each of the determinants

$$J\left[\frac{\partial f/\partial x_1,\ \partial f/\partial x_2,\ \ldots,\ \partial f/\partial x_i}{x_1,\ x_2,\ \ldots,\ x_i}\right] > 0$$

[6] Heat Transfer Analyst, Flight Propulsion Laboratory Department, General Electric Company, Cincinnati, Ohio. Assoc. Mem. ASME.

[7] "Compact Heat Exchangers," by W. M. Kays and A. L. London, The National Press, Palo Alto, Calif., 1955.

[8] "Selection of Optimum Configurations for a Heat Exchanger With One Dominating Film Resistance," by E. R. G. Eckert and T. F. Irvine, ASME Paper No. 56—S-20.

for i is equal in turn to $1, 2, \ldots, n$; when evaluated at the critical point.

Where the n x's are not all independent, but related by m constraints, the situation is far less simple. The best the authors can offer at present is: Form the $n \times n$ matrix

$$\left\| \frac{\partial^2 f}{\partial x_i \partial x_k} + \sum_{j=1}^{m} \mu_j \frac{\partial^2 \psi_j}{\partial x_i \partial x_k} \right\|$$

where the m μ's are arbitrary multipliers, different in general from the Lagrangian multipliers. Each of the determinants of order higher than $n - m$ formed from the leading minors of this matrix can be set separately equal to zero, and from the simultaneous solution of these equations, the values of the m μ's can be found. The remaining matrix of order $n - m$ must be positive definite for a minimum, rather than a saddle point, to obtain. This test is admittedly an impractical one unless the original matrix contains a large fraction of null elements.

The authors heartily concur in Mr. Daskin's desire to have an analytic function which relates the performance characteristics of variety of heat-transfer surfaces, but we politely decline the challenge.

Tests of Free Convection in a Partially Enclosed Space Between Two Heated Vertical Plates

By ROBERT SIEGEL[1] and R. H. NORRIS[2]

The free-convection heat-transfer coefficients were measured, as a function of spacing, for two parallel, electrically heated, vertical plates. The top of the rectangular space between the plates was left open, and during most of the tests the bottom and sides were closed. The heat input per unit area was substantially uniform, and the Grashof number, based on plate height, was of the order of 10^{10}. The results demonstrate how the surface-temperature rise increases, or the local Nusselt number decreases, as either of the cross-section dimensions of the free-convection space is reduced. This decrease in Nusselt number is less rapid than was expected from purely two-dimensional flow considerations. This is ascribed to the effect of an asymmetrical flow pattern observed in the space. In some cases, a periodic reversal of this asymmetrical flow also was observed. When the space between plates was opened sufficiently at the bottom, the results were reasonably consistent with the correlation proposed by Jakob for a single vertical plate with a turbulent boundary layer; these results were almost independent of spacing between the heated surfaces down to the minimum spacing tested.

Nomenclature

The following nomenclature is used in the paper:

b = breadth (width) of heated plate (see Fig. 1), ft
c_p = specific heat at constant pressure, Btu/(lb deg F)
g = acceleration of gravity, ft/hr²
h = local coefficient of heat transfer = $q''/(t_w - t_\infty)$, Btu/(hr ft² deg F)
h^* = value of h at $x = 0.81L$ for two heated plates, with space closed at bottom and sides, and $b = 0.75L$, $s = 0.28L$
k = thermal conductivity, Btu/(hr ft deg F)
L = total height of plate, ft
q'' = heat flux per unit area at wall, Btu/(hr ft²)
s = spacing between a pair of plates (see Fig. 1), ft
t = temperature, deg F
V = average velocity of horizontal air stream across the top of the plates, fpm
x = co-ordinate of height along flat plate, measured from bottom edge, ft
x_f = distance of bottom edge of plate above the floor, ft
β = volumetric coefficient of expansion, deg F⁻¹
δ^* = displacement thickness of the boundary layer, ft
μ = coefficient of viscosity, lb/(hr ft)
ν = kinematic viscosity, μ/ρ, sq ft per hr
ρ = density, pcf

Subscripts

∞ = properties of air at ambient or room temperature (see Reduction of Data for discussion)
L = location at, or value for, total height of plate
w = location at wall (surface of heated plate)

Dimensionless Groups

N_{Gr} = local Grashof number, $g\beta(t_w - t_\infty)x^3/\nu^2$
$N_{Gr,L}$ = total Grashof number based on total plate height, $g\beta(t_{w,L} - t_\infty)L^3/\nu^2$
N_{Nu} = local Nusselt number, hx/k
N_{Pr} = Prandtl number, $c_p\mu/k$

Introduction

The main purpose of this experimental investigation was to determine the free-convection heat transfer as a function of spacing between two parallel electrically heated vertical plates having a uniform heat input per unit area. The space between the plates was partially confined in most of the tests by closing the bottom and both sides with thermal insulation as shown in Fig. 1.

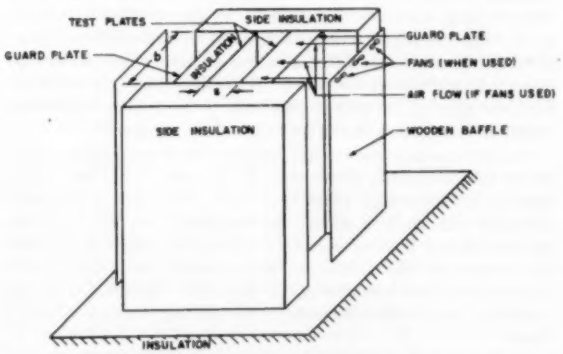


Fig. 1 Schematic Arrangement of Test Surfaces for Most of Tests

To the authors' knowledge, previous systematic data on the effect of spacing between a pair of heated plates have been available (1, 2, 3)[3] only for the case where the space between the plates was open at the bottom. A consideration of the free-convection boundary layers on plates with the space closed at the bottom led to the expectation that below some critical spacing the rising boundary layers would interfere with each other and hence

[1] Aeronautical Research Scientist, National Advisory Committee for Aeronautics, Cleveland, Ohio; formerly, General Engineering Laboratory, General Electric Company, Schenectady, N. Y. Assoc. Mem. ASME.

[2] Consulting Engineer, General Engineering Laboratory, General Electric Company, Schenectady, N. Y. Mem. ASME.

Contributed by the Heat Transfer Division and presented at a joint session with the Gas Turbine Power Division at the Semi-Annual Meeting, Cleveland, Ohio, June 17–21, 1956, of The American Society of Mechanical Engineers.

Note: Statements and opinions advanced in papers are to be understood as individual expressions of their authors and not those of the Society. Manuscript received at ASME Headquarters, January 31, 1956. Paper No. 56—SA-5.

[3] Numbers in parentheses refer to the Bibliography at the end of the paper.

obstruct the supply of cool air from the top of the space. With further decreases in spacing the heat-transfer coefficient would be expected to decrease progressively. Owing to the difficulty of analysis for such a configuration, it was decided to resort to an experimental program to establish the magnitude of these geometrical effects.

In addition to this basic configuration, several other geometries and effects were tested. These included changing the widths of the heated surfaces, blowing a stream of air horizontally at various velocities across the top of the space, heating only one plate with the other insulated, admitting air through openings between the bottom edge of the surfaces and the floor, and opening the sides of the space.

The height of the heated surfaces was 70 in., which resulted in Grashof numbers of the order of 10^{10}, based on plate height. The boundary layers on the plates were therefore expected to be turbulent except near the bottom, and this was indeed found to be the case throughout the experiments.

Experimental Equipment

The basic configuration of the test surfaces is shown schematically in Fig. 1. To help minimize stray heat losses each of the test surfaces was backed by thermal insulation and an independently heated guard plate. The heat input to the guard plate was adjusted to keep its average temperature approximately equal to the average temperature of the test plate. Six layers of aluminum-foil insulation were provided between the test and guard plates to reduce to a very low value any heat flow resulting from the unavoidable small local inequalities of temperature between the two plates.

Each heated surface was constructed of three electrically heated panels each 70 in. high and 17.5 in. wide. These were fastened together by thin vertical aluminum moldings to form a single continuous surface about 53 in. wide. The moldings projected beyond the surface of the panel about $^1/_8$ in. The heating panels were composed of a layer of bakelite insulation 0.020 in. thick covered on each side by a sheet of aluminum 0.005 in. thick. Very thin heating wires were imbedded in the central insulation layer. The spacing of these heating wires per inch was found to be substantially the same, 2.29 per in., for all parts of the panel; the wires themselves were not always perfectly straight, however, so that the spacing between a particular pair of wires sometimes varied as much as ± 10 per cent from the average.

Calculations have shown that the edgewise thermal conductance of the thin aluminum sheets was low enough so that when considering the 70-in-high plate as a whole, the surfaces had substantially uniform heat input. Nevertheless, over the small distance from one heating wire to the next, the edgewise thermal conductance of the aluminum was sufficient to maintain substantially uniform temperature of the surface, unaffected by the occasional small nonuniformity of the spacing between adjacent wires.

During most tests the space for free convection between the plates had a horizontal width b of 53 in. parallel to the plates, which were likewise 53 in. wide. In some tests, however, two thin vertical wooden partitions were inserted to reduce the width of the free-convection space. These partitions were arranged symmetrically, so that the center of the test surface of reduced width remained at the center of the 53-in-wide heating surface.

A horizontal air flow was provided across the tops of the heated panels in certain of the tests, by three axial-flow fans as indicated in Fig. 1. Although the resulting air flow was somewhat nonuniform and rather turbulent, it served to permit evaluation of the sensitivity of the test-plate temperatures to such a horizontal air stream, with a minimum expense for provision of the flow. The average velocity of this flow was measured with a velometer.

Tests also were performed with horizontal wooden-strip turbulence promoters attached to the heated plates at vertical intervals of 6.5 in. These strips projected outward from the plates $^3/_4$ in., and were $^5/_{32}$ in. thick vertically.

Instrumentation

The total power input to the test panels was determined from voltmeter and ammeter readings with a wattmeter being used as a check. The electrical load had negligible reactance, and hence a power factor of unity.

Temperatures on the test panels were measured by "Stickon" resistance-thermometer elements (Ruge DeForest Company, type BN-1), which were cemented to the plates by a thin coating of Armstrong cement. Preliminary tests with elements placed at corresponding positions on both the exposed and insulated sides of a test panel showed good agreement with each other, and hence the elements were thereafter placed only on the insulated side so that lead wires would not interfere with the natural convection flow.

Thermometer elements were located at vertical heights above the bottom edge of the plate of 3, 7, 12, 21, 36, and 61 in. One row of elements was provided in the central portion of the test surface. In addition, at the 7, 21, and 61-in. levels, elements were also located at points 9 in. horizontally inward from each of the outer vertical edges of the panels to provide an indication of the horizontal variation of the surface temperature. No attempt was made to explore more completely the temperature distribution over the entire plate surface.

The resistance of each element was measured by a Wheatstone-bridge circuit. The temperature rise of the element due to the current through the bridge was found to be negligible during the brief interval of current flow required while balancing the bridge. The bridge balance was measured by a light-beam galvanometer. A calibration chart and an individual correction factor for each element were furnished by the manufacturer to meet the accuracy called for by government specification Mil. B5495. These were spot-checked against a temperature standard for randomly chosen sample elements and the discrepancy found negligible. Corrections were of course made for the resistance of the lead wires.

Ambient-air temperatures were measured by similar resistance elements, each suspended inside a vertical cylindrical radiation shield of aluminum foil.

The air-flow patterns in the space between the plates were explored to a limited extent by observing the flow of smoke from cigarettes attached to long probes inserted from above.

Transient air-temperature fluctuations in the convection space were measured by thermistors of pin-head size with leads of fine wire soldered across the tips of two needles. These were mounted on long rods used as adjustable probes. The time constant of these thermistor elements was a small fraction of a second. The transient temperatures were recorded on a photoelectric recorder.

The major limitation to accuracy of the test results was the existence of small temperature fluctuations, as explained later. These fluctuations were greater for some configurations than for others.

Reduction of Data

Location of Ambient-Temperature Measurements. The ambient temperature was generally about five deg higher at the level of the top edge of the test plates than at the floor. The choice of the vertical location for the ambient temperature used in evaluating the heat-transfer coefficient of the plate therefore requires explanation

In the tests of a single isolated plate, the ambient temperature was taken at the same height as the surface-temperature measurement being considered.

In tests of parallel plates having their bottom edges raised off the floor, the ambient temperature was taken close to the floor, since all or most of the ambient air entering the region between the plates was sucked in from that level. This floor ambient was used even when the space between the heated surfaces was open for entry of air at the sides since it was believed that the air flow still came mostly from the bottom.

In tests with the space between the plates closed both at the bottom and at the sides, the ambient temperature was taken at the level of the top edges of the plates, since the air entering the convection space came from approximately that location.

Physical Properties of Air. The physical properties of air used to evaluate Nusselt, Grashof, and Prandtl numbers were taken at a temperature halfway between the average plate temperature and the suitable ambient-air temperature. The values chosen were consistent with reference (4). Since the local temperatures over the plate in any given test generally varied from the average by the order of only 10 deg F or less, no attempt was made to adjust the air properties for this variation.

Stray Heat Losses. The Nusselt-number results were not corrected for stray heat losses through the insulation nor for the heat losses by radiation through the top or side openings. The Nusselt numbers obtained are therefore slightly higher than values representing pure free convection alone. The heat loss through the insulation was estimated to be about 5 per cent of the total input, and the radiation loss another few per cent, the latter varying of course with the distance between the plates. The radiation emissivity of the plates was about 0.05 since the material was uncoated aluminum.

Dimensionless Correlation of Results. The results were first correlated in dimensionless form by plotting, for each configuration, the local Nusselt number hx/k, as a function of the product of the local Grashof number, $g\beta(t_w - t_\infty)x^3/\nu^2$, times Prandtl number, $c_p\mu/k$.

One reason for using this correlation was to compare the results with previously proposed correlations of this form for a single vertical plate. Another reason was that, except for the closely spaced configurations discussed later, the results obtained for one particular set of gas properties, temperature difference, and plate height, could be presumed applicable to any different set yielding the same Grashof-Prandtl-number product, without additional tests of the individual effects of these parameters.

The local Grashof number is identical with the product

$$\left[\frac{x}{L}\right]^3\left[\frac{g\beta(t_w - t_\infty)L^3}{\nu^2}\right]$$

The right-hand factor is the "total" Grashof number, based on the total height L. In the present tests the variation in local Grashof number is almost entirely due to variation of the factor (x/L); the total Grashof number varied only slightly throughout the tests.

The slope of the correlation curves, in logarithmic co-ordinates, obtained by variation of the factor (x/L) in these tests, had practically the same value as the slope obtained by previous investigators for a single vertical plate when varying factors in the total Grashof number. Hence, in the present tests of parallel plates, if the total Grashof number were varied by changing the temperature difference while (x/L) remained fixed, it would be expected that the Nusselt number would follow almost the same curve as that obtained by varying the (x/L)-factor. Since the Nusselt number depends on a small fractional power of Grashof number, a relatively large change in $(t_w - t_\infty)$ would be required to determine the magnitude of any small difference in the slope of the curves for these two cases. It was not practicable to undertake tests for this purpose with much smaller or larger values of $(t_w - t_\infty)$; the accuracy would have become poor for very small values, and a much higher surface temperature would have damaged the electrical insulation in the heated panels. A few results obtained with a 20 per cent reduction in $(t_w - t_\infty)$ did not show any significant deviation from the curve obtained by variation of (x/L).

Grashof-Number Effects for Close Spacings. The displacement thickness of a turbulent boundary layer on a single isolated vertical plate cooled by free convection has been found theoretically (6) to be given by the relation

$$\delta^*/x = 2.04(N_{Gr})^{-1/5}(N_{Pr})^{-0.52}$$

In the derivation the free-convection heat-transfer correlation proposed by Jakob (7) was employed, which applies to the range of Grashof numbers considered here.

When two heated plates are brought close together, the ratio s/L of spacing to height at which the boundary layers of the two plates will touch, or interfere by a definite amount, may therefore be expected to be proportional to $(N_{Gr})^{-1/5}$ for a given value of N_{Pr}. Since the 1/6 power is such a low power, it would take a relatively large change in $(N_{Gr})_L$ between two tests to establish the accuracy limitations of this hypothesis, and no attempt to do so was made in the present program.

The results reported for the effect of s/L therefore apply only to the particular values of $(N_{Gr})_L$ representative of the tests; namely, about 1.8 (10^{10}). In the absence of data to the contrary it is suggested that heat-transfer correlations for other values of $(N_{Gr})_L$ in the turbulent range may be predicted by replacing s/L in the present correlation by the quantity

$$(s/L)(1.8 \times 10^{10}/N_{Gr,L})^{1/5}$$

Basis of Comparison for Different Configurations. To show the effect of changing a configuration by variation of one of its dimensions, it was desirable to cross-plot the results in another form. For this purpose the ordinate chosen is the ratio of the local heat-transfer coefficient h (for any particular configuration considered) to the local heat-transfer coefficient h^* for a particular configuration chosen as the standard of reference. The rather arbitrary choice for this reference configuration is the one with two heated plates, 53 in. wide, spaced 20 in. apart, with the bottom and sides closed, and the fans not running; h^* was evaluated at the height $x = 0.81L$, for which $N_{Gr}N_{Pr}$ was approximately 10^{10}.

One set of curves on the cross plots has been taken for a fixed value of the product of local Grashof and Prandtl numbers equal to 1.0×10^{10}. This use of a common, fixed, Grashof-Prandtl-number product, makes the (h/h^*) ratio closely represent the relative heat rate q'' for a common fixed temperature rise $(t_w - t_\infty)$ rather than the relative temperature rise for a fixed heat rate. A second set of curves has been cross-plotted for a Grashof-Prandtl-number product of 5×10^9, and a third for 1.85×10^9. These represent results at levels on the plate of 30 and 10 per cent, respectively, of the total height L, for the same temperature rise for which the first set of curves represents a level of 81 per cent. The reference heat-transfer coefficient h^*, utilized in the second and third sets, has been retained at the same value as in the first set.

Results

Surface Temperatures for Single Plate. The observed variation of surface temperature vertically along a single, isolated uniformly heated plate is shown in Fig. 2. For about the bottom 12 in. the temperature rises in proportion to about the 0.2 power of the height but remains nearly constant above the 12-in. level. This behavior is consistent with theoretical expectations when the

boundary layer is assumed laminar up to the 12-in. level, and turbulent above that. The theoretical slope of 1/5 is a consequence of the fact that the case considered here is one of uniform heat flux rather than the more customary case of uniform temperature, (see reference 5 or the more recent reference 8).

Effects of Configuration Changes on Nusselt Numbers. The correlations of the test results in the form of Nusselt number as a function of the local Grashof-Prandtl-number product are presented in Figs. 3 to 7.

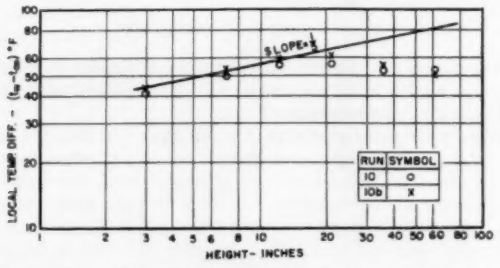


FIG. 2 LOCAL TEMPERATURE RISE OF SURFACE OF A SINGLE 70-IN-HIGH UNIFORMLY HEATED VERTICAL PLATE AS A FUNCTION OF HEIGHT ABOVE BOTTOM EDGE, FOR $N_{Gr,L}N_{Pr} = 1.5 \times 10^{10}$

Fig. 3 presents both theoretical and experimental results for the single isolated plate, and also the experimental results for a pair of parallel plates, with the space between them open at both sides, at the bottom, and at the top.

The theoretical correlation shown in Fig. 3 for a single plate with a laminar boundary layer, and uniform heat flux q'' is that derived in reference (5)

$$hx/k = 0.453(N_{Gr}N_{Pr})^{1/4} \quad [1]$$

This is about 9 per cent above the corresponding correlation for uniform wall temperature from reference (7).

The correlation proposed by Jakob (7) for a turbulent layer on a vertical plate, shown in Fig. 3, is

$$hx/k = 0.129(N_{Gr}N_{Pr})^{1/3} \quad [2]$$

This relation also applies for the case of uniform heat flux, as shown in reference (5).

For a single plate, the test results in Fig. 3 for the turbulent range are 5 to 10 per cent above Equation [2]. This agreement is quite satisfactory, the discrepancy being attributed to stray heat losses. In the laminar range, which extends over about the lower 15 per cent of the plate, the discrepancy is somewhat greater and is not fully accounted for. However, this is probably not too significant when making relative comparisons of

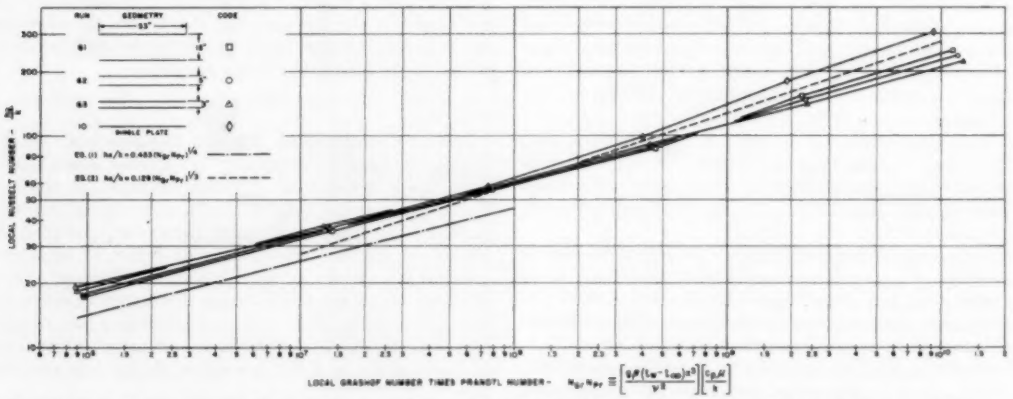


FIG. 3 TEST RESULTS FOR A PAIR OF VERTICAL HEATED PLATES WITH SPACE BETWEEN THEM OPEN AT SIDES AND BOTTOM COMPARED WITH TEST RESULTS AND PREVIOUS CORRELATIONS FOR A SINGLE PLATE ALONE
(Plate height 70 in ; plate width 53 in.; clearance above floor 5 in.; average $N_{Gr,L}N_{Pr} = 1.8 \times 10^{10}$.)

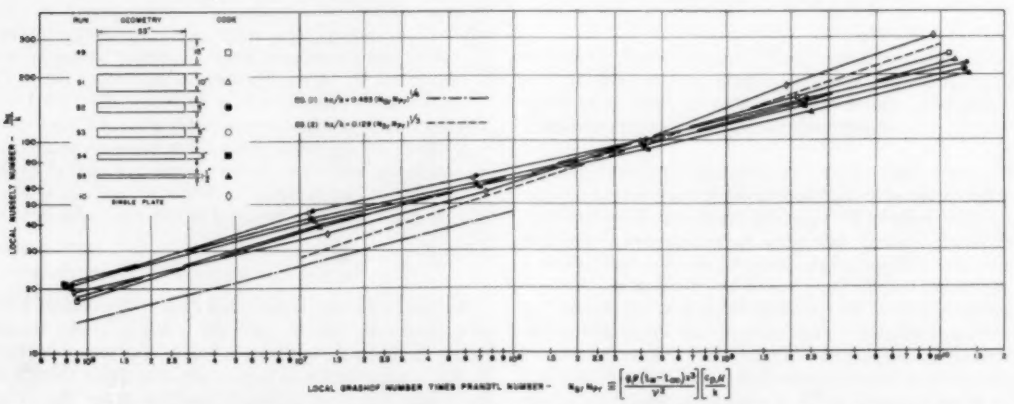


FIG. 4 TEST RESULTS FOR SAME CONFIGURATIONS AND CONDITIONS AS IN FIG. 3 EXCEPT THAT SPACE BETWEEN PAIR OF PLATES WAS CLOSED AT SIDES BY INSULATION EXTENDING DOWN TO FLOOR

the heat transfer from various configurations which have been tested.

The results in Fig. 3 for parallel heated plates are generally slightly below those for the single plate, but are relatively insensitive to variation of the spacing between plates over the range tested. Carpenter and Wassell (2) observed a large decrease in heat transfer, but this was at a spacing of only 1 per cent of plate height.

Fig. 4 presents the test results for the same configurations as Fig. 3 except that the space between the plates was closed at both sides by insulated panels which extended from the top all the way to the floor. Thus a vertical rectangular duct was formed open to the room at both bottom and top. A comparison of Figs. 4 and 3 shows that closing the sides has only a very slight effect over the upper portion of the plates. The most significant effect is the appreciable increase in the heat-transfer coefficient near the bottom of the plates at the smaller spacings due presumably to the greater "chimney" effect with the sides closed. It is noted that decreasing the spacing tends to increase the laminar characteristics of the flow as evidenced by the slopes of the correlation curves.

the downward flow of fresh, cooler air in the middle portion of the space. At spacings as small as 10 per cent of the plate height, h is still about 80 per cent of the value it has for the largest spacing, whereas the boundary layer-displacement thickness at the top of each plate has been calculated from Equation [15] in reference (6) to be 5 per cent of the plate height, if unaffected by the presence of any adjacent plate. Thus for the 10 per cent spacing the two theoretical boundary layers would just touch each other and tend to block seriously the downward flow of cooling air. This would be expected to decrease h to much less than the 80 per cent value just mentioned. A reason for the better-than-expected heat transfer actually observed under these conditions will be considered later.

The effect of adding the previously described turbulence promoters is also shown in Fig. 6(b) (cross-plotted from curves of the same type as Fig. 6(a) but not presented in this paper). It is evident that the effect of the promoters is relatively small. For the larger spacings between plates they increase h by a few per cent on the upper portion of the plates, but for the smaller spacings, and near the bottom part of the plate for all spacings, they appreciably decrease h.

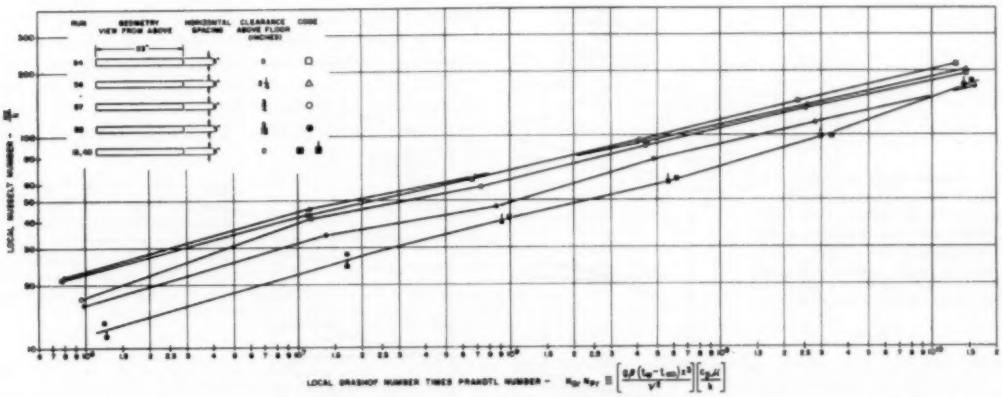


FIG. 5(a) TEST RESULTS FOR EFFECT OF VARIATION OF CLEARANCE OF BOTTOM EDGE OF PLATES ABOVE FLOOR, FOR ONE CONFIGURATION OF FIG. 4; NAMELY, A PLATE SPACING OF 3 IN.

Test results for successive decreases in the clearance opening between the bottom edges of the plates and the floor are presented in Fig. 5(a) for one of the configurations considered in Fig. 4. The particular horizontal spacing between the plates chosen for this purpose was 3 in. (0.043 L). The restrictions to the flow of air into the bottom of the vertical test duct were not thin sharp-edged plates, but flat ducts 14.5 in. long, since the insulated space on the back of each heated plate was 14.5 in. thick (see Fig. 1). The cross plot, Fig. 5(b), shows that for the particular test conditions, the heat-transfer coefficient was relatively insensitive to the clearance except for clearance below 1 per cent of the plate height.

The effect of varying the horizontal spacing s between the plates (for the widest plates tested, and with the bottom and both sides closed) is shown in Fig. 6. For smooth plates, without the fans operating, the heat-transfer coefficient is relatively insensitive to the spacing for spacings above 20 per cent of the plate height. Although it decreases at spacings smaller than this, the decrease is not as severe as had been expected from estimates of the free-convection boundary-layer thickness on individual plates and from considerations of the interference of these layers with

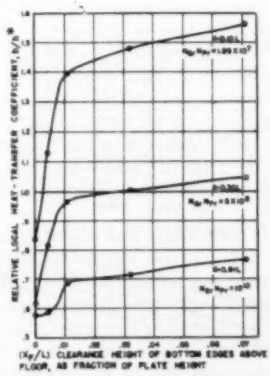


FIG. 5(b) CROSS PLOT FROM FIG. 5(a) OF RELATIVE HEAT-TRANSFER COEFFICIENT AS A FUNCTION OF CLEARANCE OF BOTTOM EDGE OF PLATES ABOVE FLOOR FOR PLATE SPACING OF 3 IN.

(h^* is value of h from Fig. 6(a) for $x/L = 0.81$, $b/L = 0.75$, $s/L = 0.28$, zero clearance at bottom and $N_{Gr}N_{Pr} = 10^{10}$.)

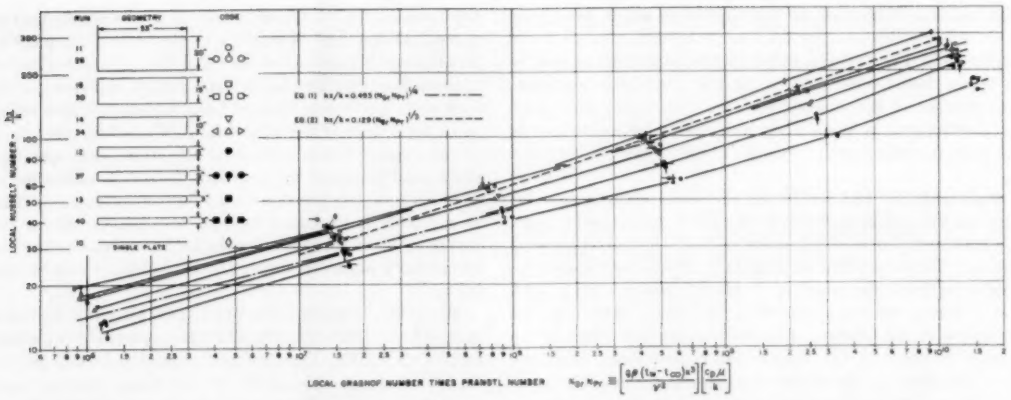


FIG. 6(*a*) TEST RESULTS FOR PAIR OF UNIFORMLY HEATED VERTICAL PLATES WITH SPACE BETWEEN THEM CLOSED AT BOTTOM AS WELL AS AT SIDES; ALSO TEST RESULTS AND PREVIOUS CORRELATIONS FOR A SINGLE PLATE ALONE

(Horizontal "ticks" on left or right sides of test points refer to temperatures 9 in. from left or right sides, respectively, of heated plate; other test points were taken near middle of plate; plate height 70 in.; plate width 53 in.; average $N_{Gr,L} N_{Pr} = 1.8 \times 10^{10}$.)

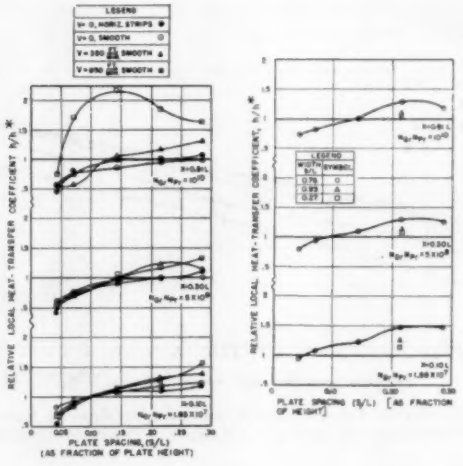


FIG. 6(*b*) CROSS PLOT OF EFFECT OF PLATE SPACING ON RELATIVE HEAT-TRANSFER COEFFICIENTS, FOR CONFIGURATIONS AND CONDITIONS CONSIDERED IN FIG. 6(*a*)

(Both plates heated. Results with addition of horizontal strips, or of a horizontal air stream across the top, are also shown.)

FIG. 7(*b*) CROSS PLOT FROM FIG. 7(*a*) OF EFFECT OF PLATE SPACING ON RELATIVE HEAT-TRANSFER COEFFICIENTS FOR VERTICAL PLATES, WITH ONLY ONE PLATE HEATED, AND WITH SPACE BETWEEN PLATES CLOSED AT SIDES AND BOTTOM

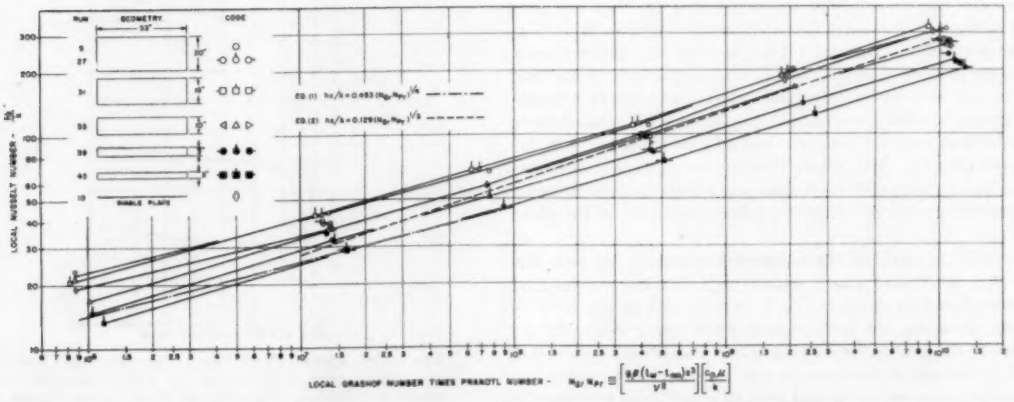


FIG. 7(*a*) TEST RESULTS FOR SAME CONFIGURATIONS AS IN FIG. 6, EXCEPT THAT ONLY ONE PLATE WAS HEATED INSTEAD OF BOTH

The effect of providing a horizontal air flow across the tops of the plates is also shown in Fig. 6(*b*). Near the tops of the plates, this air flow is seen to improve the heat transfer as would be expected. The results plotted for $x = 0.81L$ apply only to the heated plate nearest the source of the horizontal air stream. The values of h were considerably greater on the other heated plate at this level owing to the more direct impingement of the air stream. On the lower portions of the heated plate, the effect of the horizontal flow is relatively small, and at close spacings actually decreases h, as a result perhaps of its interference with flow patterns described later. The effect in this region was found to be the same for both plates.

Fig. 7 shows results for the same configurations as Fig. 6 but with only one plate heated and the other insulated.

Fig. 7(*b*) is cross-plotted from Fig. 7(*a*) and from similar curves which are not shown. A comparison with Fig. 6(*b*) shows that the effect of plate spacing is not greatly different, but that with only one plate heated, the heat-transfer coefficient is 10 to 25 per cent higher than for both plates heated. This is consistent qualitatively with the greater space provided for downward flow by the absence of an upward flowing boundary layer on the unheated plate. These values of h are also seen to be higher than the experimental results for the single plate [compare Figs. 7(*a*) and 3].

The effect of decreasing the width b of the convection space for each of two different, fixed, spacings between the pair of heated plates is shown in Fig. 8. The results show a significant decrease in the heat-transfer coefficient, even for widths considerably greater than the spacing between plates.

Flow Patterns and Flow Pulsations. Figs. 9 and 10 pertain to the results of explorations of the flow patterns in the space between plates, with the bottom and sides closed. Fig. 9 represents schematically the pattern which might be expected from two-dimensional considerations of the boundary-layer formation on a single heated vertical flat plate. This was substantially the pattern observed for a pair of heated plates when the spacing between the plates was large; i.e., 28 per cent of the plate height.

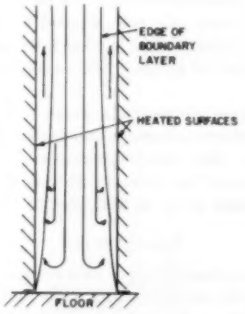


FIG. 9 TWO-DIMENSIONAL SCHEMATIC DIAGRAM INDICATING FLOW PATTERN WHICH MIGHT BE EXPECTED BETWEEN A PAIR OF VERTICAL HEATED PLATES WITH SPACE BETWEEN THEM CLOSED AT SIDES AND BOTTOM

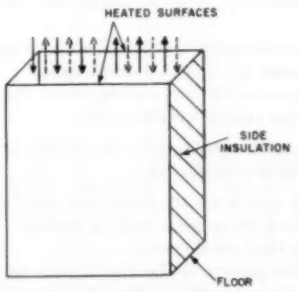


FIG. 10 THREE-DIMENSIONAL SCHEMATIC DIAGRAM INDICATING ASYMMETRICAL FLOW DISTRIBUTION OBSERVED AT TOP OF A DUCT CLOSED AT BOTTOM AND SIDES, WHEN SPACING BETWEEN HEATED PLATES WAS 21 PER CENT OR LESS OF HEIGHT

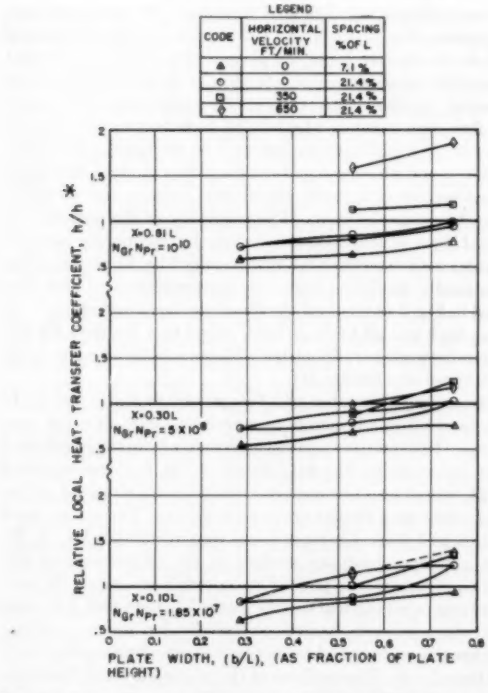


FIG. 8 CROSS PLOT OF EFFECT OF WIDTH OF TEST DUCT (CLOSED AT SIDES AND BOTTOM) ON RELATIVE HEAT-TRANSFER COEFFICIENTS, FOR TWO DIFFERENT PLATE SPACINGS, WITH AND WITHOUT HORIZONTAL AIR STREAM ACROSS THE TOP; BOTH PLATES HEATED; AVERAGE $N_{Gr,L}N_{Pr} = 1.8 \times 10^{10}$

When the spacing was reduced to $0.21L$, the width being $0.75L$, the interference of the upward flowing boundary layers with the downward flow, required to feed those layers, apparently modified the flow patterns. The cross section occupied by the downward flow no longer extended over nearly the full width of the plate in a symmetric manner, but was confined to about one half the width, as indicated by the solid arrows in Fig. 10, with the other half width having only upward flow. This asymmetric pattern is presumably the reason why the heat-transfer coefficient fell off less rapidly, with decreasing spacing, than had been expected.

With spacings of $0.21L$ or less, and widths of $0.75L$, a periodic fluctuation in the flow velocity was observed in the region between the plates. When the spacing was reduced to $0.07L$, the fluctuation amplitude was so large that actual reversals in flow direction took place periodically as indicated by the dotted arrows in Fig. 10. Similar reversals occurred at a spacing of $0.04L$. Records of the temperature of this fluctuating air flow were consistent with flow observations made with smoke; the air temperature near the top, halfway between the heated plates and near one side, was close to room temperature at the same time that the air temperature at a similar point near the other side was much

warmer. A moment later the temperature at the first point would rise, while the temperature at the second point would drop to about room temperature. The average period of the flow reversals was about 20 sec, but with considerable variation from cycle to cycle.

When the width of the test duct was reduced to $0.28L$, with plate spacing at $0.21L$ the pattern represented by Fig. 10 was no longer evident. The temperature of the air halfway between the heated plates and near one side showed irregular fluctuations representative of turbulent flow, but at intervals of from 5 to 40 sec abruptly dropped to a value close to room temperature for 1 to 2 sec duration.

For a duct with an intermediate width of $0.53L$, and a plate spacing of $0.21L$ sometimes one and sometimes the other of the foregoing types of flow and temperature fluctuations was observed. This resulted in different values of heat-transfer coefficients in successive tests, as indicated in Fig. 8.

Conclusions

The following conclusions apply to a vertical duct of rectangular cross section with the two wider walls uniformly heated and having a Grashof-Prandtl-number product of about 1.8×10^{10}, based on total height of the plates:

1 When the duct was open at the bottom the local heat-transfer coefficients were affected only slightly by the following:

(*a*) The presence or absence of vertical enclosing walls at the sides.

(*b*) The spacing between the heated walls, down to the minimum spacing tested (2 per cent of duct height).

(*c*) The clearance of the bottom edge above the floor, down to a clearance of 1 per cent of the duct height.

2 When the duct was closed at the bottom and at the sides, the following effects were observed:

(*a*) The local heat-transfer coefficients were appreciably affected over most of the ranges tested, by both duct width and the spacing between the heated plates.

(*b*) The air-flow pattern in the duct became asymmetric, with a tendency to pulsate or reverse in direction for certain geometries; namely, ducts of large width and relatively close spacing between the heated plates.

(*c*) Provision of a horizontal air stream across the top of the duct can cause a small decrease in the heat-transfer coefficients of the lower portions of the plates. It tends to increase the coefficients for the upper portions except when the spacing between plates is quite small.

(*d*) Provision of horizontal turbulence-promoter strips, on the heated plates, results in no significant change in the heat-transfer coefficient except at the closest plate spacing tested, where it results in a considerable decrease due to blocking the air flow.

(*e*) The heat-transfer coefficient for a heated plate is somewhat larger when the plate faces an unheated plate rather than a heated one.

BIBLIOGRAPHY

1 "Heat Dissipation of Parallel Plates by Free Convection," by W. Elenbaas, *Physica*, vol. 9, January, 1942, p. 1.

2 "The Loss of Heat by Natural Convection From Parallel Vertical Plates in Air," by L. G. Carpenter and H. C. Wassell, Proceedings of The Institution of Mechanical Engineers, London, England, vol. 128, November-December, 1934, pp. 439–457.

3 "Studies on Heat Transfer in Laminar Free Convection with the Zehnder-Mach Interferometer," by E. R. G. Eckert and E. E. Soehngen, Air Force Technical Report 5747, December, 1948.

4 "The NBS-NACA Tables of Thermal Properties of Gases—Tables 2.10, 2.39, and 2.42 (Dry Air)," by H. W. Woolley, F. C. Morey, and R. L. Nutall, National Bureau of Standards, 1950.

5 "Analysis of Laminar and Turbulent Free Convection From a Smooth Vertical Plate With Uniform Heat Dissipation per Unit Surface Area," by Robert Siegel, General Electric Report 54GL89, April, 1954.

6 "Analytical Investigation of Flow and Heat Transfer in Coolant Passages of Free-Convection Liquid-Cooled Turbines," by E. R. G. Eckert and T. W. Jackson, NACA RM E50D25, July, 1950.

7 "Heat Transfer," by M. Jakob, John Wiley & Sons, Inc., New York, N. Y., 1949, p. 530.

8 "Laminar Free Convection From a Vertical Plate With Uniform Surface Heat Flux," by E. M. Sparrow and J. L. Gregg, Trans. ASME, vol. 78, 1956, pp. 435–440.

Discussion

Frank Kreith.[4] The data presented by the authors represent a valuable contribution to the understanding of free convection in partially enclosed spaces. This type of information is not only important to the development of free-convection cooling systems for turbine blades and nuclear power plants, but it is also of fundamental interest.

In the following discussion some of the results presented in the paper are re-examined and compared with published information on related systems and unpublished data obtained at Lehigh University on a geometrically similar system. It is hoped that these remarks will foster a continued interest in the fundamental aspects of the problem.

An inspection of Fig. 3 reveals that the experimental data for the lower portion of a single vertical plate, where the flow is believed to be entirely laminar, fall 20 to 30 per cent above the line representing a theoretical equation predicted by Siegel, and deviate even further from theory when compared with an analysis by Sparrow and Gregg (9)[5] which gives a constant of 0.408 compared to the value of 0.453 in Equation [1]. Similar experimental results were obtained at Lehigh University with an electrically heated copper plate, 3 ft tall and 1 ft wide, in a range of Grashof numbers in which over $^2/_3$ of the plate is covered by a turbulent free-convection boundary layer. In view of the deviation of these experimental results from theoretical predictions for purely laminar flow, one wonders whether the turbulent free-convection boundary layer could have an influence on the laminar boundary layer below it and whether equations for purely laminar flow apply to free convection in systems where both laminar and turbulent boundary layers are present. The reliability of the authors' experimental data is strengthened further by comparing some of their results with experimental data obtained by Elenbaas (1) on a geometrically similar system. A rearrangement of the data presented in Fig. 3 in terms of the dimensionless numbers used by Elenbaas, hs/k and $g\beta\Delta Ts^4c_p\mu/Lkv^2$, shows that the data for free convection between parallel plates with open bottoms agree quite well with those of reference (1).

The dimensionless number $g\beta\Delta Ts^4c_p\mu/Lkv^2$ or $N_{Gr,s} \cdot N_{Pr} \cdot (s/L)$ can also be used to advantage in correlating data of other configurations. Using the best estimates possible from the small-sized graphs in the preprint, the data shown in Fig. 6(*a*) are replotted in Fig. 11, where also some experimental points obtained at Lehigh University on a similar system are shown. The system used in the tests at Lehigh University was constructed by Mr. A. W. Stubner, a former graduate student in the department of mechanical engineering, as a part of his master's program. It consists of two parallel vertical copper plates 3 ft high and 1 ft wide, each of which can be heated electrically by passing current through heating wires imbedded between sheets of mica on the back sides of the plates. The surfaces of the plates on which the heat-

[4] Associate Professor of Mechanical Engineering, Lehigh University, Bethlehem, Pa. Assoc. Mem. ASME.

[5] Numbers from 9 to 12 in parentheses refer to the Bibliography at the end of this discussion.

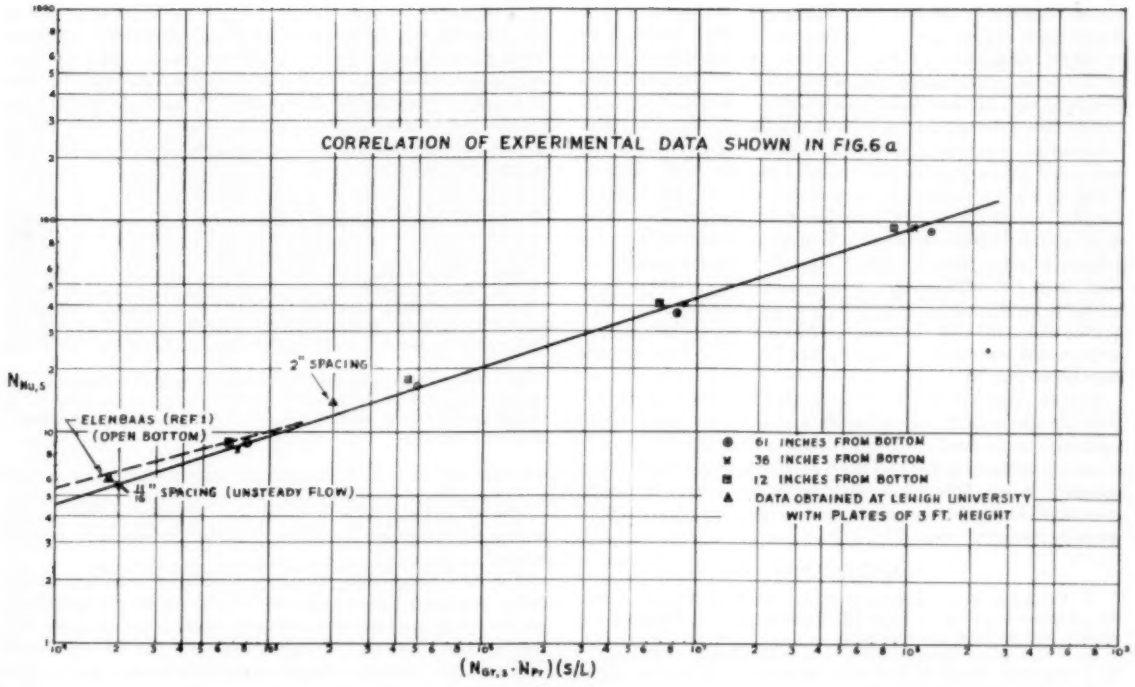


FIG. 11

ing wires are installed were covered with asbestos insulation. The heat loss through the insulation was measured as a function of the plate temperature by placing the plates face to face, heating them, and measuring their temperature and the heat dissipation at equilibrium. The plate temperatures were measured at various locations with iron-constantan thermocouples peened into the copper plates from the rear. The results of two tests with the plates 2 in. and $^{11}/_{16}$ in. apart, sides and bottom closed, are shown in Fig. 11. They are in good agreement with the data from Fig. 6(*a*), and suggest that a Grashof number based on the distance between the plates may be a significant indication of the turbulence for this system, somewhat analogous to a Reynolds number based on the hydraulic diameter or radius in forced convection. The use of the plate spacing as the significant length dimension appears to be particularly useful (1, 10) when that distance is so small that the entire space between the plates is filled with fluid influenced in some way by viscous forces.

It is of interest to note that in those systems where the bottom was open (see Figs. 3 and 4) the curves representing the Nusselt numbers for the smaller openings cross the curves representing the Nusselt numbers for the larger openings. This trend is not observed in any of the systems in which the bottom was closed. This discusser would like to ask if an explanation for these phenomena can be found from an examination of the flow in a manner similar to that employed by Lighthill (10), or Batchelor (11) in their respective analyses of free convection problems.

The description of the periodic flow fluctuations which were observed when the distance between the plates was reduced to less than 20 per cent of their height is possibly the most interesting phase of the experimental work. Lighthill (10), in his study of free convection in a tube closed at the bottom, showed theoretically how the flow pattern changes when the boundary layer becomes so thick at the top that the inflow of colder fluid is choked. His results, however, do not apply here because in a tube, symmetry of the flow is maintained even after choking, whereas in a narrow space between flat plates the flow pattern ceases to be symmetrical about the center plane (see Fig. 9) and becomes circulatory. Although the appearance of this type of motion is physically not unexpected since it allows for continued flow and effective heat transfer after the central inflow pattern becomes choked, a reason for its unsteadiness cannot be found readily. However, similar flow instabilities have been observed by Sparrow and Kaufman (12) in a narrow vertical space fed by fluid from a

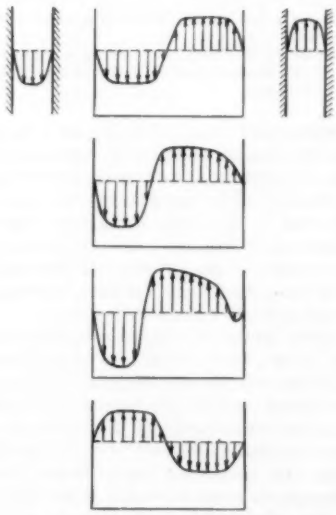


FIG. 12 SKETCHES ILLUSTRATING FLOW REVERSAL

heated reservoir at the bottom and cooled at the top, and their observations suggest an explanation for the flow reversal observed by the authors. It may be noted also that, in the system treated in the present paper, the natural tendency of heated fluid is always to rise. Hence, in a narrow channel fed by fluid from the top also the inflowing fluid is subject to a buoyant force which in the case of circulatory motion is, however, weaker than the downward forces produced by upward moving fluid. But, as shown in Fig. 12, there must be somewhere in the space between the plates some fluid which is neither moving upward nor downward because it is located at an intermediate boundary where the velocity is zero. This boundary will shift as the stationary fluid rises under the influence of buoyancy and thus reduce the area available for downflow. The shift continues until the boundary has moved so far that the inflow becomes choked. The choking action could conceivably produce a flow reversal as shown by the sketches in Fig. 12.

This discusser would like to suggest that the authors present in their closure typical samples of data showing the temperature variations in the fluid at the open end as well as the variations in surface temperature (that is of the heat-transfer coefficient) during unsteady flow. This information would be of help in further studies of flow phenomena associated with free convection in partially confined spaces.

BIBLIOGRAPHY

9 "Laminar Free Convection from a Vertical Plate with Uniform Surface Heat Flux," by E. M. Sparrow and J. L. Gregg, Trans. ASME, vol. 78, 1956, pp. 435–440.

10 "Theoretical Considerations in Free Convection in Tubes," by M. J. Lighthill, *Quarterly Journal of Mechanics and Applied Mathematics*, vol. VI, 1953, pp. 398–439.

11 "Heat Transfer by Free Convection Across a Closed Cavity Between Vertical Boundaries at Different Temperatures" by G. K. Batchelor, *Quarterly Journal of Applied Mathematics*, vol. XII, No. 3, 1954, pp. 209–233.

12 "Visual Study of Free Convection in a Narrow Vertical Enclosure," by E. M. Sparrow and S. J. Kaufman, NACA RM No. E 55L14a, Feb. 16, 1956.

LLOYD TREFETHEN.[6] It would be helpful if the authors could provide more information about the asymmetrical flow pattern of Fig. 10, and an explanation of the periodic reversal of that flow which they observed. The high density of the fluid moving down compared to that of the hotter rising fluid, together with the equality of pressure of fluid as it enters and leaves, would seem to tend toward stability of the asymmetrical flow once it has established itself in either direction. Yet the authors report the interesting fact that such a flow is not always stable.

The somewhat similar flow patterns obtained by Timo (13)[7] were stable. There, liquid metal in a narrow annulus between vertical adiabatic cylinders was frozen at the top, and contacted a pool of hot liquid metal at the bottom of the cylinders. Perhaps the rising hot streams scalloped the metallic ice in such a way as to provide a stabilizing geometry. On the other hand, Sparrow and Kaufman (14) cooled the top of water between vertical adiabatic glass plates which were open at the bottom to a heated reservoir. They observed violently unstable and irregular convective patterns.

This large-scale departure from two-dimensional flow conditions should perhaps be anticipated in some other instances of convective flow between vertical boundaries. For example, thermopane windows are frequently subjected to uneven temperatures, inside by heating currents, and outside by winds. These uneven effects can be seen on single-thickness glass windows in the winter, when the boundaries between the drop-condensation, film-condensation, freezing, and sublimation regions show large dips and peaks. Nonhorizontal isotherms on thermopane would tend to cause large-scale eddies of the air between the panes similar to those observed by Timo and by the authors. If such eddies did occur, the thermal resistance of the window might differ from the resistance predicted from a two-dimensional analysis, such as Batchelor's (15).

[6] Assistant Professor, Division of Engineering and Applied Physics, Harvard University, Cambridge, Mass. Assoc. Mem. ASME.

[7] Numbers from 13 to 15 in parentheses refer to Bibliography at the end of this discussion.

BIBLIOGRAPHY

13 "Free Convection in Narrow Vertical Sodium Annuli," by D. P. Timo, U. S. Atomic Energy Commission Technical Information Service, Oak Ridge, Tenn. Report KAPL-1082, March 5, 1954.

14 "Visual Study of Free Convection in a Narrow Vertical Enclosure," by E. M. Sparrow and S. J. Kaufman, National Advisory Committee for Aeronautics Research Memorandum E 55L14a, February 16, 1956.

15 "Heat Transfer by Free Convection Across a Closed Cavity Between Vertical Boundaries at Different Temperatures," by G. K. Batchelor, *Quarterly Journal of Applied Mechanics*, vol. XII, No. 3, 1954, pp. 209–233.

AUTHORS' CLOSURE

The authors would like to thank the discussers for their interest in and comments on the subject of free convection in confined spaces.

The data supplied by Professor Kreith for heat transfer between two heated plates are very welcome as they provide independent information which tends to confirm the authors' findings for this geometry. It was interesting that his experiments using a single plate also yielded values in the laminar region which were higher than the theoretical laminar correlations. Contrary to the statement in the discussion, the analysis of Sparrow and Gregg (reference 8 of the paper) is in good agreement for Pr = 0.7 with that of Siegel (5). For, if we use Fig. 4 of reference (8) and also the notation of that reference and note the definitions of Gr_x^* and of Nu_x, we have $Nu_x/Gr_x^{*1/5} = Nu_x/(Gr_x Nu_x)^{1/5} = 0.5$. Then, for Pr = 0.70, $Nu_x = (0.5)^{5/4}\,(Gr_x Pr)^{1/4}/(Pr)^{1/4} = 0.458\,(Gr_x Pr)^{1/4}$. This factor 0.458 agrees substantially with 0.453 in Equation [1] of the paper, which came from reference (5).

Professor Kreith comments on the crossing of the Nusselt-number curves in Figs. 3 and 4. An analysis of this system might require a consideration of the laminar to turbulent transition for free convection in a channel, which is not well understood at present. Ignoring this transition effect, however, the following explanation of the crossing of the curves seems not unreasonable. As the plates are brought closer together, the resulting increase in air temperature between the upper portions of the plates creates a better draft (chimney effect) which increases the upward velocity between the lower plates. This improves the heat transfer there, since the boundary layers are then thin enough so as to leave room for unheated ambient air between the two layers. This effect is greater when the sides are closed (Fig. 4) than when they are open (Fig. 3) as already mentioned in the paper. But between the upper portions of the plates the thermal boundary layers have merged, so closer spacing means higher air temperature at the center, which reduces the heat transfer.

With reference to Professor Trefethen's comment on the observed flow reversals, the authors are unable to supply additional information to clarify the origin of this flow configuration. Professor Kreith offers a possible explanation, and it would be desirable to have additional experimental information on the details of the flow during the oscillations to establish his hypothesis.

To better illustrate the nature of the oscillations, Figs. 13 and 14 herewith show typical recordings of air-temperature varia-

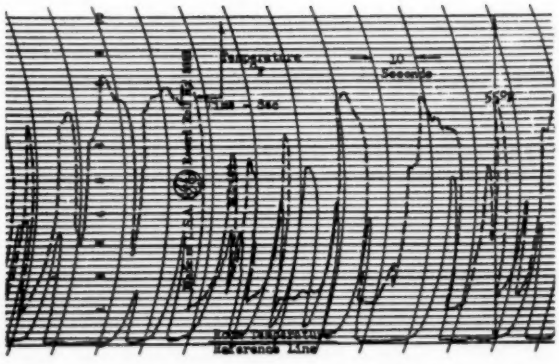


FIG. 13 AIR TEMPERATURES HALFWAY BETWEEN HEATED PLATES AND LEVEL WITH TOP EDGE

(Solid curve for point $0.06L$ from one side; dotted curve $0.03L$ from other side. $b/L = 0.75$; $s/L = 0.07$.)

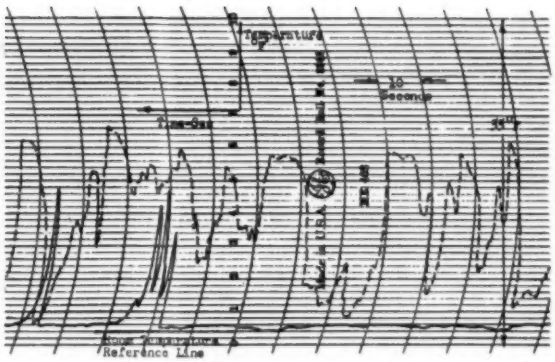


FIG. 14 AIR TEMPERATURES HALFWAY BETWEEN HEATED PLATES AND LEVEL WITH TOP EDGE

(Solid curve for point $0.03L$ from one side; dotted curve $0.03L$ from other side. $b/L = 0.53$, $s/L = 0.07$.)

tions in the space between two heated plates with the sides and bottom closed. Recordings of this nature are discussed toward the end of the paper in the section on flow pulsations. No attempt was made to record the fluctuations in plate temperature since the output of the resistance thermometers attached to the heated surfaces would not be representative of the temperature-fluctuations of the portions of the plates where these thermometers were not present, owing to differences in time constant of temperature response to changes in the convective heat-transfer coefficient.

A Mechanical Computing Device for the Analysis of One-Dimensional, Transient, Heat-Conduction Problems

By W. E. HOWLAND,[1] E. A. TRABANT,[2] AND G. A. HAWKINS,[3] LAFAYETTE, IND.

In 1953 one of the authors learned that Dr. P. Hackemann, the year before, had designed and constructed a mechanical device for use in the solution of one-dimensional, transient, heat-conduction problems (1, 2).[4] The device replaces the laborious task of line drawing required in the Binder-Schmidt graphical method by a semi-automatic machine operation. This paper covers an extension of the Hackemann idea to the solution when the thermophysical properties vary with temperature. Before introducing the suggested mechanical computer, a brief discussion of the Binder-Schmidt graphical method for solving one-dimensional, transient, heat-flow problems is given.

Illustrative Use of Binder-Schmidt Method for a Simple Heat-Transfer Problem

THE Binder-Schmidt procedure is basically a graphical solution of a finite-difference equation which replaces the basic partial differential equation used to express the transient-conduction heat transfer in the system.

For the purpose of illustration consider the partial differential equation for unidimensional flow of heat in a bar of uniform cross section and length l

$$\frac{\partial t}{\partial \tau} = \alpha \frac{\partial^2 t}{\partial x^2} \quad \ldots\ldots [1]$$

where

t = temperature in rod at any position
τ = time
α = thermal diffusivity of rod material, or
$\alpha = k/\rho c$
k = thermal conductivity of material
c = specific heat
ρ = density

For initial condition it is assumed that

$$t = t(x), \quad \tau = 0 \quad \ldots\ldots [2]$$

where $t(x)$ is an arbitrary function, and the boundary conditions are taken as

$$\left.\begin{aligned} t &= t_1 \quad \text{at} \quad x = 0 \\ t &= t_2 \quad \text{at} \quad x = l \end{aligned}\right\} \quad \ldots\ldots [3]$$

To solve Equation [1] numerically it is usual to replace the derivatives with respect to time or distance by finite differences. If

$$t_n(x) \text{ and } t_{n+1}(x)$$

are the temperature distributions at the beginning and end of the nth time interval (τ_n, τ_{n+1}) of length $\Delta\tau$, then

$$\frac{t_{n+1} - t_n}{\Delta\tau} = \frac{k}{\rho c} \frac{d^2(t_n)}{dx^2} \quad \ldots\ldots [4]$$

With the following definition

$$\frac{1}{\beta^2} = \frac{\rho c}{k \Delta\tau} \quad \ldots\ldots [5]$$

Equation [4] may be written

$$\frac{d^2(t_n)}{dx^2} = \frac{1}{\beta^2}(t_{n+1} - t_n) \quad \ldots\ldots [6]$$

If $t_{m,\,n}$ and $t_{m+1,\,n}$ represent the temperature at the beginning and end of the mth distance interval (x_m, x_{m+1}) of length Δx and at the beginning of the nth time interval, then Equation [6] may be written as

$$t_{m,\,n+1} - t_{m,\,n} = \frac{\beta^2}{(\Delta x)^2}[t_{m+1,\,n} - 2t_{m,\,n} + t_{m-1,\,n}] \quad \ldots\ldots [7]$$

Equation [7] is a formula of recursion which indicates that the temperature at a position $x = m\Delta x$ may be calculated for the instant $\tau = (n+1)\Delta\tau$ if the distribution in the neighborhood of x is known at $\tau = n\Delta\tau$.

A graphical solution of Equation [7] may be accomplished in the following manner: In Fig. 1 the curve a-c-e-g-i-k represents the arbitrary initial condition, Equation [2], and the points a and k correspond to the boundary conditions, Equation [3]. The bar has been divided into 10 increments Δx, as shown.

If a straight line is drawn from e to g, it is evident that

$$\overline{ff'} = \frac{1}{2}(t_{6,0} + t_{4,0}) - t_{5,0}$$

or

$$\overline{ff'} = \frac{1}{2}(t_{6,0} - 2t_{5,0} + t_{4,0}) \quad \ldots\ldots [8]$$

Comparison of Equations [7] and [8] shows that if the quantity $\overline{ff'}$ is multiplied by the factor $(\beta/\Delta x)^2$ the result will be the increase of temperature at the position $x = 5\Delta x$ in a time interval $\Delta\tau$. If the points b, d, h, and j are found in a manner similar to f then these points represent the temperature at the time $\tau = \Delta\tau$.

---

[1] Professor of Sanitary Engineering, Purdue University.
[2] Associate Professor of Engineering Sciences, Purdue University.
[3] Dean of Engineering, Purdue University. Mem. ASME.
[4] Numbers in parentheses refer to the Bibliography at the end of the paper.

Contributed by the Heat Transfer Division and presented at the Semi-Annual Meeting, Cleveland, Ohio, June 17–21, 1956, of THE AMERICAN SOCIETY OF MECHANICAL ENGINEERS.

NOTE: Statements and opinions advanced in papers are to be understood as individual expressions of their authors and not those of the Society. Manuscript received at ASME Headquarters, January 31, 1956. Paper No. 56—SA-27.

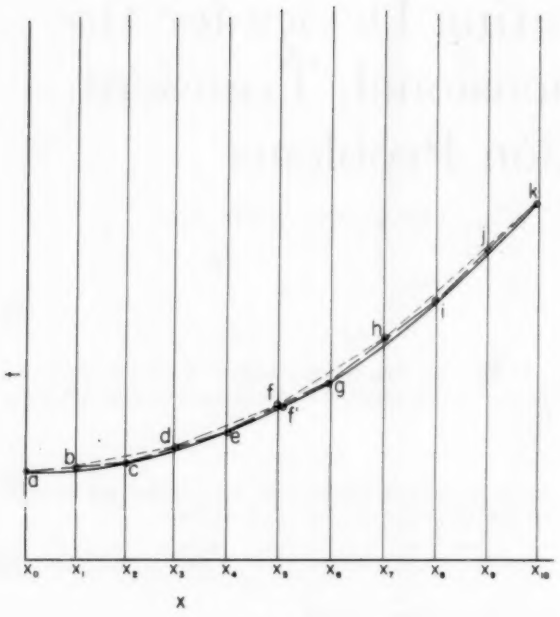


FIG. 1

A continuation of this construction will produce lines of temperatures in equal steps of $\Delta\tau$.

It is observed that, if the graphical procedure is carried far enough, the end result will be a straight line connecting the points a and k. This agrees with the analytic steady-state solution.

Moreover, since it is always possible to choose values of Δx and $\Delta\tau$ such that

$$\beta^2/(\Delta x^2) = \frac{1}{2} \quad \text{or} \quad \Delta\tau = (\Delta x)^2/2\alpha \ldots\ldots\ldots [9]$$

Equation [7] may be written in the simplified form

$$t_{m,\,n+1} = \frac{1}{2}\,(t_{m+1,\,n} + t_{m-1,\,n}) \ldots\ldots\ldots\ldots [10]$$

HACKEMANN IDEA

In this example it is easy to see that an elastic string could be made to take the successive positions of the temperature-distribution curve if it were operated in the following manner: First, let it be stretched to take the positions *a-c-e-g-i-k*. Then let it be held at the positions *a-b-d-f-h-j-k* and released at points *c, e, g, i*. Then let it be clamped in these new positions where the string crosses the vertical lines above the points *c, e, g, i*, and so on. The string will always be held at a and k but alternately on the solid vertical lines. A mechanism that alternately would clamp the string along every other vertical line and then on the others is not difficult to imagine. Such a device has been constructed by Dr. Hackemann.

FLOW OF HEAT THROUGH COMPOSITE WALLS

The possible use of an elastic-string computer in the study of temperature distributions in composite walls provides an interesting illustrative example.

There are two principal difficulties in the way of applying the elastic-string method to this problem: (*a*) When the string is stretched across the interface, it naturally will be a straight, unbroken line as the support at the interface is removed. This straightness of line would imply that the temperature gradient is a constant across the interface, which it certainly is not when the k-values of the two materials are different. (*b*) When the k-values of the two materials are different, the calculated values of $\Delta\tau$ would be different in the two sections as shown by Equation [9] providing Δx is the same throughout. Under these conditions, the string would have to move more rapidly through part of its length than in the other. This would require two strings and is impractical. The remedy for the second difficulty is obviously to change the value of Δx in the different parts of the path to make $\Delta\tau$ constant everywhere and this can be done readily. The remedy for the first difficulty is to replace each section of wall by a different one of different thickness, but of the same conductivity as every other and of resistance to heat-flow equivalent to that of the wall replaced and of equivalent total heat capacity. This also may be done as will now be shown.

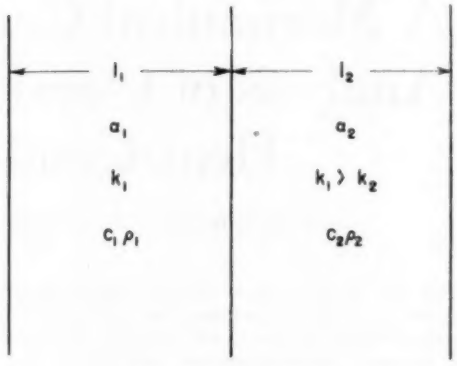


FIG. 2 ACTUAL WALLS

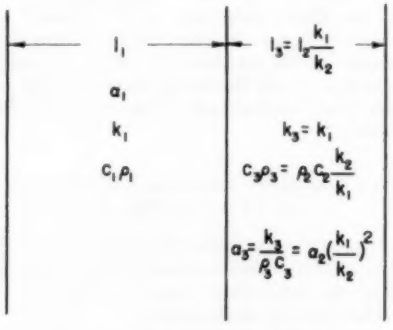


FIG. 3 SUBSTITUTE WALLS

Replace the actual wall shown in Fig. 2 by the nearly equivalent wall shown in Fig. 3. The difference is that the material designated by subscript 2 has been replaced by another one whose conductivity value k_3 is equal to k_1, and whose thickness has been changed so that the resistance to the steady flow of heat is the same as that of the material replaced. This means that the thickness of the new material is

$$l_3 = l_2 \frac{k_3}{k_2} = \frac{l_2 k_1}{k_2} \ldots\ldots\ldots\ldots [11]$$

The new wall also must have the same heat capacity as the wall replaced. This means that the product $c_3\rho_3$, of the material of the

new wall must be changed in the ratio of the lengths or the conductivities. Therefore

$$c_1\rho_1 = c_2\rho_2 \frac{k_2}{k_3} = c_3\rho_3 \frac{k_3}{k_1} \quad \ldots\ldots\ldots\ldots [12]$$

For example, if the length has been shortened as shown, i.e., if $l_2 > l_3$ then more heat capacity must be crowded into the same space so that $c_3\rho_3 > c_2\rho_2$.

The two walls shown in Fig. 3 are now exactly equivalent to those shown in Fig. 2. Since the conductivity is the same throughout, a constant difference in temperature between the outsides of the combined wall of Fig. 3 will produce a straight unbroken temperature line for steady flow. Furthermore, for unsteady flow the slope of the temperature gradient at the interface will be continuous.

It now appears that the elastic-string analogy to the flow of heat through walls of different material will present no serious difficulties at the interface. Across the interface, when not held fixed at the interface, the elastic string will be pulled straight, implying that the temperature gradient is continuous at the interface—which is true when the conductivities are arbitrarily adjusted to be the same in both wall materials, by Equations [11] and [12].

It is recalled that the time intervals $\Delta\tau$ must be the same for all materials since one string must be used for the entire gradient. This is accomplished by keeping $\Delta\tau$ the same throughout. When α_3 is different from α than Δx_3 must differ from Δx_1 in accordance with Equation [9]. In the illustrative example presented, since

$$c_3\rho_3 l_3 = c_2\rho_2 l_2$$

$$k_3 = k_1 \quad \text{and} \quad \frac{k_2}{l_2} = \frac{k_3}{l_3}$$

then

$$c_3\rho_3 = c_2\rho_2 \frac{l_2}{l_3} = c_2\rho_2 \frac{k_2}{k_3} = c_2\rho_2 \frac{k_2}{k_1}$$

and

$$\frac{\Delta x_3}{\Delta x_1} = \frac{(2\alpha_3\Delta\tau)^{1/2}}{(2\alpha_1\Delta\tau)^{1/2}} = \sqrt{\left(\frac{\alpha_3}{\alpha_1}\right)} = \sqrt{\left(\frac{k_1\rho_1 c_1}{k_3\rho_3 c_3}\right)} \ldots\ldots [13]$$

With composite sections an additional difficulty may often be encountered. After an appropriate choice of Δx_1 has been made and Δx_3 has been calculated, it may be found to be not an aliquot part of the wall thickness l_3. In other words, the total equivalent thickness l_3, of the second part of the wall may not be an integral multiple of the calculated value of Δx_3. Suppose that l_3 divided by Δx_3 comes out to be n, an integral number, plus a fraction f. Choose n as the number of sections and divide the length l_3 into this number of sections. Place clamping bars of width

$$(l_3 - n\Delta x_3)/n = f\Delta x_3/n \ldots\ldots\ldots\ldots [14]$$

at the boundaries of these sections each l_3/n wide. The center line of a clamping bar should coincide with the boundaries of the sections. At the interface and end place clamping bars of one half this width. If a drafting-board procedure is to be used, allow a gap of these magnitudes to exist between the ends of the chords as shown in Fig. 4. The chords will have a horizontal projected width of exactly Δx_3 as calculated.

If the cross-sectional areas of the two sections are different as well as the corresponding thermal constants, then the second section is replaced by a third whose resistance to heat flow and whose total heat capacity are the same as those of the section replaced

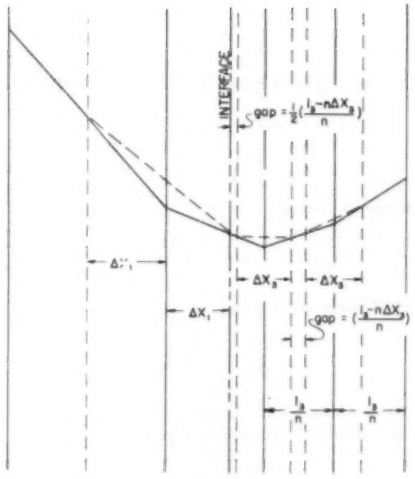


FIG. 4

but, in addition, whose product of thermal conductivity and area is the same as that of the first section. This is

$$\left.\begin{aligned} A_1k_1 &= A_3k_3 \\ \frac{A_3k_3}{l_3} &= \frac{A_3k_3}{l_2} \end{aligned}\right\} \ldots\ldots\ldots\ldots [15]$$

But

so

$$l_3/l_2 = A_2k_3/A_3k_3$$

or

$$k_3/k_2 = (l_3/l_2)(A_2/A_3) \ldots\ldots\ldots\ldots [16]$$

and

$$\left.\begin{aligned} A_3l_3c_3\rho_3 &= A_2l_2c_2\rho_2 \\ A_2l_2/A_3l_3 &= c_3\rho_3/c_2\rho_2 \end{aligned}\right\} \ldots\ldots\ldots\ldots [17]$$

or

Now

$$\alpha = k/\rho c$$

so

$$\frac{\alpha_3}{\alpha_2} = \frac{k_3}{k_2}\frac{c_2\rho_2}{c_3\rho_3} = \left(\frac{l_3}{l_2}\frac{A_2}{A_3}\right)\left(\frac{A_3}{A_2}\frac{l_3}{l_2}\right) = \left(\frac{l_3}{l_2}\right)^2 \ldots\ldots [18]$$

Discussion of Method for Finding Equations for Spacing of Lines

For radial heat conduction in cylinders, the differential equation is

$$\frac{\partial t}{\partial\tau} = \alpha\left(\frac{\partial^2 t}{\partial r^2} + \frac{1}{r}\frac{\partial t}{\partial r}\right) \ldots\ldots\ldots\ldots [19]$$

By the change of variable

$$x = \ln r \ldots\ldots\ldots\ldots [20]$$

Equation [19] may be written in the form

$$\frac{\partial t}{\partial \tau} = \frac{\alpha}{r^2}\frac{\partial^2 t}{\partial x^2} \quad [21]$$

For radial heat flow in spherical shells the differential equation is

$$\frac{\partial t}{\partial \tau} = \alpha\left(\frac{\partial^2 t}{\partial r^2} + \frac{2}{r}\frac{\partial t}{\partial r}\right) \quad [22]$$

By the change of variable

$$x = -\frac{1}{r} \quad [23]$$

Equation [22] may be written

$$\frac{\partial t}{\partial \tau} = \frac{\alpha}{r^4}\frac{\partial^2 t}{\partial x^2} \quad [24]$$

Thus it is seen that both Equations [19] and [21] may be put into a form similar to Equation [1]; namely

$$\frac{\partial t}{\partial \tau} = \Phi\frac{\partial^2 t}{\partial x^2} \quad [25]$$

where Φ is not a constant as in Equation [1] but a function of the variable x.

Now as in Equation [9] for the case of the rod, so here for the cylinder let

$$\Phi\Delta\tau/(\Delta x)^2 = \frac{1}{\alpha}$$

Since for the cylinder $\Phi = \alpha/r^2$ then

$$\frac{\alpha}{r^2}\frac{\Delta\tau}{(\Delta x)^2} = \frac{1}{2}$$

or

$$\Delta\tau = \frac{(\Delta r)^2}{2\alpha} = \frac{(e^x\Delta x)^2}{2\alpha} = \frac{(\Delta x)^2}{2\Phi} \quad [26]$$

Similarly as in the case of the rod and cylinder, for the sphere let

$$\frac{\alpha}{r^4}\frac{\Delta\tau}{(\Delta x)^2} = \frac{1}{2}$$

or

$$\Delta\tau = \frac{(\Delta r)^2}{2\alpha} = \frac{(\Delta x)^2}{2\alpha x^4} = \frac{(\Delta x)^2}{2\Phi} \quad [27]$$

since for the sphere

$$\Phi = \alpha/r^4$$

For these two cases the Binder-Schmidt method simply requires that the differential equation be put in the form of Equation [25] where Φ is not necessarily a constant but may be some function of x, and then choose Δx, the space between the parallel lines, in accordance with the equation

$$\Delta x = \sqrt{[2\Phi(\Delta\tau)]} \quad [28]$$

The question naturally arises: "Could not Φ be other functions of x besides those here dealt with and also could it not be a function also of t?" The latter case would result in curved lines rather than in parallel straight lines. If Φ were allowed to vary with other functions of x and also with t, the generality of the method would be increased to include practical problems in which the thermal properties themselves are affected by both the variation of x and of t, if not, indeed, of time itself.

General Method of Choosing Spacing of Lines or Bars on Computing Device

In the expression

$$\Delta x = \sqrt{(2\Phi\Delta\tau)}$$

assume that $\Delta\tau$ is a constant and the appropriate value of

$$\Delta x/\sqrt{[\Phi(x, t)]}$$

can be found by

$$\sqrt{(2\Delta\tau)} = \int_{x_a}^{x_b}\frac{dx}{\sqrt{[\Phi(x, t)]}} \quad [29]$$

x_a and x_b are the corresponding x-values on two successive lines Δx apart horizontally at a particular value of t. Now the value of

$$z = \int_{x_a}^{x_b}\frac{dx}{\sqrt{[\Phi(x, t)]}} \quad [30]$$

one way or another, can be evaluated for particular values of t. Here x is a general value of x and a value of t will be chosen which will give the maximum value of spacing Δx. The curve of z versus x can be plotted for all values of x along the path of heat flow, say z vertically and x horizontally. Then the ordinate of this curve can be divided into sections of equal length

$$\Delta z = \sqrt{(2\Delta\tau)} \quad [31]$$

The values of x corresponding to the extremities of these Δz-sections will be the points on the successive lines or bars the correct Δx distance apart. Then the values of Δx will be computed at the middle of the spaces so determined, but at various values of t, from the equation

$$\Delta x = \sqrt{[2\Phi(x, t)\Delta\tau]}$$

The extremities of these spaces Δx_a, Δx_b, Δx_c, etc., would then be plotted at various values of t, the temperature. It is noted that the change in Φ alters the incremental slab thickness. However, compensation is not made for variation in over-all thickness owing to a temperature-dependent conductivity. The graph or computing device might then appear as shown in Fig. 5. The shaded

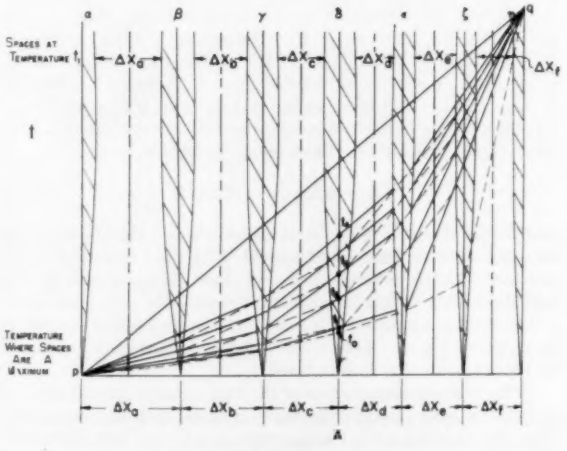


Fig. 5 Maximum Spaces x

portion of the sketch, if it is to be considered as a Hackemann mechanism, may be thought of as movable bars or clamps: α, β, γ, etc., as shown. The initial temperature distribution is indicated by the initial position of the elastic string, which is represented by the lowest inclined and broken dotted line shown in Fig. 5. It is held in position by all the bars α, β, γ, and so on, which press the elastic string down firmly on a horizontal board. The ends of the elastic string are always held firmly at points p and q. Then the alternate bars β, δ, and ζ are raised and the elastic string stretches to the position shown by the next higher inclined broken line (shown as a continuous line). Then the bars β, δ, ζ are lowered to press against the string and the bars γ and ϵ are raised and the string takes the position shown by the next higher inclined broken line (shown as a dotted line). The successive positions of the string indicate successive temperature distributions at successive equal intervals of time. Thus t_0, t_1, t_2, etc., are the temperatures at section A at successive equal increments of time, $\Delta\tau$.

All the lines and curves could, of course, have been constructed with a pencil on a drafting board, most of them with a straight edge, without the aid of an elastic string or clamping mechanism, but such a device, if perfected, might prove useful in expediting the procedure, especially when a very large number of sections Δx and time intervals $\Delta\tau$ are to be employed.

It is of interest to apply the method to Equation [1] and the examples given in reference (2).

For Equation [1], $\Phi(x, t) = \alpha = \text{const}$, and using Equations [30] and [31]

$$\Delta z = \int_{x_a}^{x_b} \frac{dx}{\sqrt{\alpha}} = \frac{x_b - x_a}{\sqrt{\alpha}} = \frac{\Delta x}{\sqrt{\alpha}} = \sqrt{(2\Delta\tau)}$$

or

$$\Delta x = \sqrt{(2\alpha\Delta\tau)}$$

which confirms the result, Equation [9].

For Equation [21]

$$\Phi(x, t) = \frac{\alpha}{r^2} = \frac{\alpha}{e^{2z}}$$

and Equations [30] and [31] give

$$\Delta z = \int_{x_a}^{x_b} \frac{dr}{\sqrt{(\alpha/e^{2z})}} = \frac{e^{x_b} - e^{x_a}}{\sqrt{\alpha}} = \frac{r_b - r_a}{\sqrt{\alpha}}$$

or

$$\Delta r = \sqrt{(2\alpha\Delta\tau)}$$

which confirms the previous analysis as shown in Equation [26]. In order to obtain the required equal increments of r, plot e^z and choose appropriate increments of x therefrom.

Equation [24] may be handled in a similar manner.

The heat-flow equation may be written in the more general form

$$\frac{\partial\left[\Psi(r, t)\,\dfrac{\partial t}{\partial r}\right]}{\partial r} = \frac{\Psi(r, t)}{\alpha}\,\frac{\partial t}{\partial \tau} \quad \ldots\ldots\ldots\ldots [32]$$

By the transformation of variable defined by

$$\frac{\partial x}{\partial r} = \frac{1}{\Psi(r, t)} \quad \ldots\ldots\ldots\ldots [33]$$

Equation [32] may be changed in the following manner:

Now

$$\frac{\partial t}{\partial r} = \frac{\partial t}{\partial x}\,\frac{\partial x}{\partial r} = \frac{1}{\Psi(r, t)}\,\frac{\partial t}{\partial x}$$

so

$$\frac{\partial}{\partial r}\left[\Psi(r, t)\,\frac{\partial t}{\partial r}\right] = \frac{\partial}{\partial r}\left[\frac{\Psi(r, t)\,\dfrac{dt}{\partial x}}{\Psi(r, t)}\right] = \left[\frac{\partial}{\partial x}\left(\frac{\partial t}{\partial x}\right)\right]\frac{\partial x}{\partial r}$$

$$= \frac{\partial^2 t}{\partial x^2}\,\frac{1}{\Psi(r, t)}$$

Hence in terms of the new variable x Equation [32] takes the form

$$\frac{\partial t}{\partial \tau} = \frac{\alpha}{[\Psi(r, t)]^2}\,\frac{\partial^2 t}{\partial x^2} \quad \ldots\ldots\ldots\ldots [34]$$

This shows that the general Equation [32] may be put in the form of Equation [25] and may be solved by the described computing device or the corresponding graphical procedure. In this case

$$z = \int_{x_a}^{x_b} \frac{\Psi(r, t)dx}{\sqrt{\alpha}} = \int_{z_a}^{z_b} \frac{dr}{\sqrt{\alpha}} \quad \ldots\ldots\ldots [35]$$

If α is a constant then

$$z = \frac{1}{\sqrt{\alpha}}\,\Delta r \quad \ldots\ldots\ldots\ldots [36]$$

and equal increments of z correspond with equal increments of r.

BIBLIOGRAPHY

1 Based on a discussion with Dr. P. Hackemann in London, England, in 1953.

2 "A Mechanical Apparatus for the Approximate Solution of One-Dimensional Heat Conduction Problems," by P. Hackemann, paper presented at Symposium at the R.A.E., Farnborough, July, 1954, Ministry of Supply, Heat Transfer Panel, England.

3 "Differenzenverfahren zur Losung von Differentialgleichungen," by Ernst Schmidt, *Forschung VDI*, Bd/Heft 5, September–October, 1942, p. 177.

4 "The Solution of Transient Heat Conduction Problems by Finite Differences," by G. A. Hawkins and J. T. Agnew, Engineering Bulletin, Purdue University, Research Series No. 98.

5 "Heat Transfer," by M. Jacob, John Wiley & Sons, Inc., New York, N. Y., vol. 1, 1949.

Discussion

MERL BAKER.[6] The use of the elastic-string analogy as the basis of a mechanical computing device for the analysis of one-dimensional transient heat-conduction problems is an ingenious one. The diversity of the Binder-Schmidt method has long been recognized, and when restricted by proper boundary conditions may be useful in solving almost any one-dimensional transient heat-flow problem. The disadvantage has been the labor required to achieve the desired results by graphical construction. With a suitable computing device such as proposed in this paper, the method of solution should have far reaching significance.

The writer has had a particular interest in developing suitable boundary conditions permitting application of the Binder-Schmidt method to evaluate the heat flow through massive building structures which are intermittently heated and cooled. Considerable progress has been made, with the chief objection being the manpower required to complete the construction. This is par-

[6] The Kentucky Research Foundation, Lexington, Ky.

ticularly objectionable when composite walls are considered. The computing device used to analyze this particular problem would be invaluable in determining the heating and cooling loads under transient conditions.

G. M. Dusinberre.[6] The bibliography given by the authors might be supplemented by the following:

"Applied Mathematics in Chemical Engineering," by T. K. Sherwood and C. E. Reed, McGraw-Hill Book Company, Inc., New York, N. Y., 1939.

"Application of Schmidt's Graphical Method to Cases Involving Diffusivity as a Variable," by A. H. Brown, U. S. Department of Agriculture Report ED-6-12-12, June 14, 1946.

"Numerical Analysis of Heat Flow," by G. M. Dusinberre, McGraw-Hill Book Company, Inc., New York, N. Y., 1949.

"A New Electrical Analog Method for the Solution of Transient Heat-Conduction Problems," by G. Liebmann, Trans. ASME, vol. 78, 1956, pp. 655–665.

It is of interest that Sherwood and Reed suggested that variable diffusivity might be handled by a simple transformation of co-ordinates, while Brown showed that this could not be done. Now the authors have solved the problem by a double set of transformed co-ordinates.

But it is the fate of many problem-solving techniques to remain merely academic curiosities. A conspicuous example is the "bubble analogy" which, nevertheless, someone is always trying to revive as a practical method for heat-source problems.

If the present contribution—regarded as an *engineering* technique—is to escape this fate, then the authors should show the field in which it is useful and, within that field, the economic advantage over competitive methods.

Obviously the present method is limited to one-dimensional systems. Practically, it is limited to a small number of nodes or reference points. Further, in the writer's experience, any graphical method is intolerably messy for cyclic temperature changes. The problem of a variable surface coefficient occurs as often as not; the authors do not show how to deal with this.

The field of application is thus considerably restricted. Even within this narrow field the writer feels that the method would hardly be competitive with Liebmann's analog—a tool of much greater general utility—or with numerical methods of solution.

If this estimate is unduly pessimistic, the authors are urged to refute it by examples drawn from engineering practice.

M. J. Goglia.[7] The authors have presented an extension of Hackemann's elastic-string analog computer to the one-dimensional heat-conduction problem. There is no need for justifying the effort being spent in the development of such devices—they often serve as the vehicle by means of which solutions are arrived at to problems not tractable by analytic means alone.

In this regard, two questions naturally present themselves; the one is related to the one-dimensional treatment and the second to the extension of the method discussed to two-dimensional considerations. Have the authors explored (*a*) the manner of introducing contact resistances as conditions one would be faced with if application of this device were made, for example, to laminated transformer cores, and (*b*) the procedures to be employed in the use of this device if the boundary conditions are other than the constant temperatures used for the illustrative examples in the paper? Have the authors considered the feasibility of using a two-dimensional network of elastic strings in the extension of their analog computer to two-dimensional heat-conduction problems?

Victor Paschkis.[8] Starting from what is introduced as the "Hackemann mechanism" the authors present a number of interesting extensions of the Binder-Schmidt graphical method for solution of transient heat-conduction problems. The main complaint against this method has always been the length required to carry it out. If the Hackemann mechanism can shorten the time required in a really significant manner, the method could be much more useful. Did the authors actually work with the device; and do they have any estimate regarding the time saved? How critical is the mechanical execution of the device? For example, the movable bars or clamps (Fig. 5) (*a*) will have a finite width, while they ideally should have zero width; (*b*) because of play in the joints, they may not always land on the same line, but only within a band. What magnitude error will be introduced by such unavoidable mechanical limitations?

Authors' Closure

The computing device described has not been constructed. However, the original Hackemann mechanical computer was constructed by him and proved successful in its operation. All operations of the envisioned computer in the paper have been carried out on a drafting board. For the problems considered, the error was never more than five per cent, when compared with existing analytic solutions. The problem of contact resistance and the two-dimensional extension of the computer had not been considered at the time of presentation of the paper, but will be given consideration. It is thought that mechanical devices are possible which will simulate other than constant-temperature boundary conditions.

[6] Professor of Mechanical Engineering, Department of Mechanical Engineering, Pennsylvania State University, University Park, Pa. Fellow ASME.

[7] Regents' Professor of Mechanical Engineering, Georgia Institute of Technology, Atlanta, Ga. Mem. ASME.

[8] Technical Director, Heat & Mass Flow Analyzer Laboratory, Columbia University, New York, N. Y. Mem. ASME.

A Comparison of Refrigerants When Used in Vapor Compression Cycles Over an Extended Temperature Range

By J. P. BARGER,[1] W. M. ROHSENOW,[2] AND K. M. TREADWELL,[3] CAMBRIDGE, MASS.

An important current engineering problem involves providing apparatus to meet the cooling requirements of supersonic aircraft. Because the time is already passing when simple air cycles can render adequate performance due to their inability to extend over a sufficient temperature range, vapor compression cycles are being considered. However, the maximum temperature range over which a single refrigerant can operate successfully in a vapor cycle is also relatively limited. Since the total required temperature range increases rapidly with the aircraft flight Mach number, series combination or "stacking" of vapor cycles leading to binary and higher order cycles may become necessary. This paper gives methods for selecting the refrigerant, or series combination of refrigerants, which exhibits optimum thermodynamic performance when employed in vapor cycles operating between arbitrarily selected temperatures.

Introduction

THE nature of the problem of cooling supersonic aircraft is now becoming clear. Heat will be transferred from sources inside the aircraft at various prescribed temperatures to a sink outside the aircraft at a higher temperature. Fig. 1 presents the air stagnation temperature versus flight Mach number relationship for standard NACA air. This stagnation temperature is the temperature of the sink to which the heat must be rejected from the aircraft. Even if ram air-expansion turbines, not popular with aircraft designers because of their large size and weight, can be employed, the sink temperatures at Mach number greater than 3 are still much higher than those which the refrigeration engineer normally encounters.

The heat-source temperatures depend on the function of the aircraft. In general, the entire aircraft will not be cooled; parts of it will be allowed to approach stagnation temperature. The temperature of three main areas of the aircraft will be reduced:

1 Certain parts of future hydrocarbon-burning engines will require cooling to keep metal temperatures below 2000 R.

2 The temperature of electronic equipment of present design will be kept below approximately 700 R.

3 Cabins of manned aircraft will be kept at temperatures around 510–530 R.

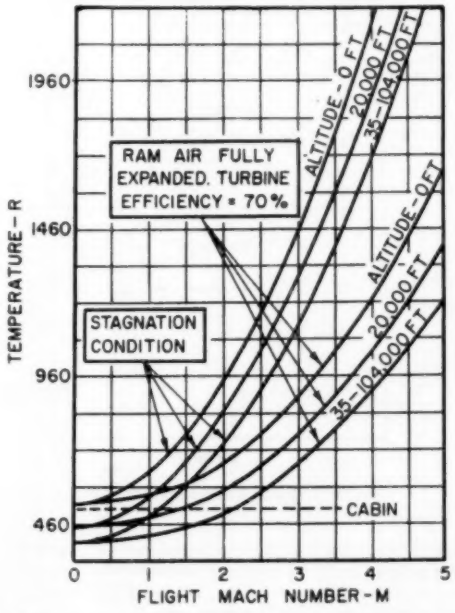


FIG. 1 STAGNATION AND RE-EXPANDED AIR TEMPERATURE VERSUS MACH NUMBER

At any given Mach number, then, the temperature range between source and sink can be determined by comparing one of the foregoing source temperatures with the stagnation temperature corresponding to the Mach number read from Fig. 1.

The following sections evaluate and compare the performance of vapor cycles using commercially available refrigerants over the required temperature ranges.

Single Vapor Cycles

Cycles. The classical refrigeration cycle of Fig. 2 incorporates the standard condenser, constant-enthalpy expansion valve, evaporator, and compressor in sequence. Friction-pressure drops have been neglected. Heat is received in the evaporator from an air stream or another condensing refrigerant and is delivered from the condenser to another air stream, or to another evaporating refrigerant.

A standard definition of cycle c.o.p. is used, which is as follows

$$\text{c.o.p.} = \frac{Q \text{ evaporator}}{W_x \text{ compressor}} = \frac{h_{R4} - h_{R3}}{h_{R1} - h_{R4}} \quad \ldots\ldots\ldots [1]$$

Carnot cycles with evaporator and condenser temperatures equal to those of the actual cycles show the thermodynamic limit of performance for the actual cycles. The solid process line of

[1] Instructor, Department of Mechanical Engineering, Massachusetts Institute of Technology. Assoc. Mem. ASME.

[2] Associate Professor, Department of Mechanical Engineering, Massachusetts Institute of Technology. Mem. ASME.

[3] Department of Mechanical Engineering, Massachusetts Institute of Technology.

Contributed by the Heat Transfer Division and presented at a joint session with the Aviation Division at the Semi-Annual Meeting, Cleveland, Ohio, June 17–21, 1956, of THE AMERICAN SOCIETY OF MECHANICAL ENGINEERS.

NOTE: Statements and opinions advanced in papers are to be understood as individual expressions of their authors and not those of the Society. Manuscript received at ASME Headquarters, January 19, 1956. Paper No. 56—SA-6.

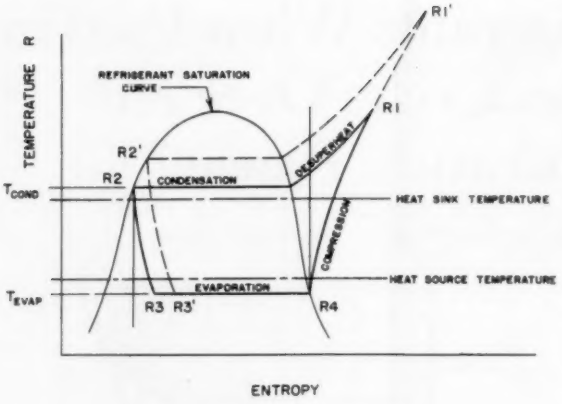


FIG. 2 ACTUAL VAPOR CYCLE

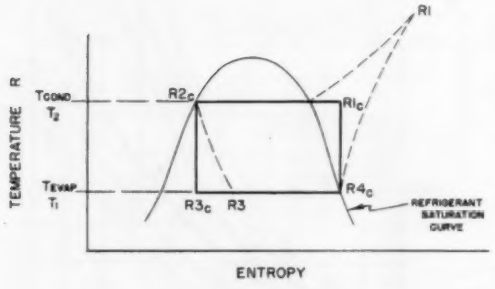


FIG. 3 CARNOT REFRIGERATION CYCLE

Fig. 3 represents such a Carnot cycle. For this cycle the compression process becomes reversible and adiabatic, eliminating the actual compressor's irreversibility; the refrigerant expands at constant entropy rather than constant enthalpy; and the irreversibility associated with the desuperheating process is eliminated. The Carnot refrigeration c.o.p. reduces to

$$\text{c.o.p.} = \frac{T_1}{T_2 - T_1} \quad [2]$$

where T_1 is the temperature at which heat is received in the evaporator, and T_2 is the temperature at which heat is rejected in the condenser.

Refrigerant Limitations. The same Carnot-cycle limitation on the coefficient of performance between two widely separated but fixed temperatures applies whether one or many cycles in series are employed to transfer heat between the temperatures. Favoring the use of the smallest number of cycles, each with the widest temperature spread, is the increased mechanical complexity of binary and higher-order arrangements. On the other hand, the c.o.p. of each single cycle decreases faster than its Carnot c.o.p. as the temperature spread between its evaporator and condenser increases because the proportion of irreversible to reversible processes in the cycle rises. In determining the minimum possible number of cycles which can be used between widely separated temperatures, the maximum temperature spread for application of each refrigerant must be evaluated.

The maximum practical temperature spread through which an actual refrigerant may be used is a function not only of temperature considerations such as the freezing point but also of pressure and specific-volume limitations.[4] These restrictions divide into two groups, those which form the upper limit of condensing temperature and those which form the lower limit of evaporating temperature.

As regards the upper limit, the critical temperature, while not an absolute barrier to raising the upper temperature of vapor compression cycles, certainly represents a practical limit. In Fig. 4 a dashed constant-enthalpy line representing a throttling expansion from the critical point shows how little enthalpy change remains to be accomplished in the evaporator. Such an expansion always results in a vanishingly small c.o.p. except for extremely small evaporator-condenser temperature differences. Supercritical condensing processes are also less efficient. Using a constant-temperature heat sink, it is possible to produce subcooled condenser exit states by raising the condensing pressure. However, no advantage results from this measure because:

[4] Table 1 lists properties of commercially available refrigerants pertinent to the following discussion on limitations on temperature spread.

TABLE 1 REFRIGERANT PROPERTIES

Refrigerant	Freezing pt.— deg F, 14.7 psia	Critical temp., deg F	Critical press, psia	Minimum table value— Temp. sat. liquid, deg F	Minimum table value— Press. sat. liquid, psia	Maximum table value— Temp. sat. vapor, deg F	Maximum table value— Press. sat. vapor, psia
Mercury	....	3002.	51400.	220.0	0.007	1390.0	1100
Water	32.0	705.4	3206.	32.0	0.09	705.4	3206.2
Carbon tet.	−9.4	541.	661.	Unstable at high temperatures			
Ethyl ether	....	552.1	380.8	....	....	....	....
Trichloroethylene	....	520.0	728.0	....	....	....	....
Dichloroethylene	....	470.0	795.0	....	....	....	....
Methylene chloride	−142.	421.0	670.0	10.0	1.38	140.0	26.79
Methyl formate	−147.5	418.0	607.0	0.0	1.50	140.0	38.41
Freon 113	−31.	417.4	495.0	−30.0	0.30	200.0	54.66
Freon 11	−168.0	388.4	635.0	−40.0	0.74	388.4	635.0
Ethyl chloride	....	369.0	764.0	....	....	....	....
Ethylamine	−115.	362.0	815.0	−58.0	0.35	113.0	41.49
Freon 21	−221.	353.3	750.0	−40.0	1.36	160.0	100.6
Sulphur dioxide	−98.	314.8	1141.5	−40.0	3.14	140.0	158.6
Methylamine	−134.	314.0	1082.0	−58.0	1.32	113.0	98.76
Butane	−211.	306.0	550.1	−30.0	3.4	180.0	160.0
Freon 114	−137.	294.3	474.0	−80.0	0.46	140.0	84.8
Methyl chloride	−153.	289.4	969.0	−80.0	1.95	170.0	283.9
Isobutane	−229.	272.7	537.0	−20.0	7.50	180.0	210.0
Ammonia	−107.9	271.2	1651.0	−107.8	0.88	125.0	307.8
Freon 12	−247.	232.7	582.1	−155.0	0.12	232.7	582.1
Carrene 7	−247.0	221.1	631.0	−40.0	10.84	140.0	263.5
Freon 22	−256.	204.8	716.0	−155.0	0.20	160.0	448.0
Propane	−310.	202.0	661.5	−75.0	6.37	140.0	305.0
Propylene	−301.	196.5	667.2	−50.0	16.	90.0	195.
Nitrous oxide	−152.	96.5	1050.0	−127.0	14.7	96.5	1050.0
Ethane	−278.	90.1	708.3	−148.0	7.62	90.1	708.3
Carbon dioxide	−109.	87.8	1071.0	−147.0	2.14	87.8	1071.0
Freon 13	−296.	83.8	589.0	−200.	0.43	83.8	579.0
Ethylene	−272.	48.8	731.8	−176.8	6.75	48.8	731.8
Freon 14	−312.	−49.9	542.4	−250.	1.1	−49.9	542.4
Methane	−297.	−115.8	673.0	−260.0	15.0	−115.8	673.0
Air	....	−221.0	547.0	....	....	....	....

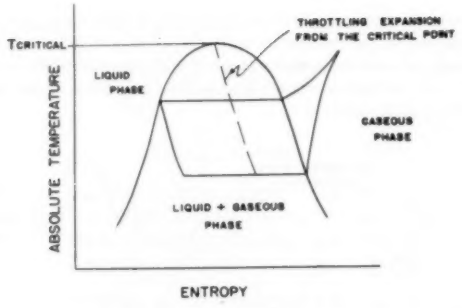


FIG. 4 *T-S* DIAGRAM FOR PURE SUBSTANCES

(*a*) Evaporator heat transfer increases only minutely as a result of the small enthalpy decrease of the prethrottling state.

(*b*) With a higher condensing pressure, a higher pressure ratio across the vapor compressor leads to considerably increased compressor work.

Between identical source and identical sink temperatures, the combination of items (*a*) and (*b*) always results in a lower c.o.p. than accompanies vapor cycles where the preexpansion condition is on the saturated-liquid line, therefore precluding consideration of any condensing process ending in the subcooled liquid state, including supercritical condensing processes.

In considering the effect of pressure on the maximum condensing temperature, the designer of practical systems will be forced to choose a pressure limit where the size and weight of the system necessary to contain the high pressure begins to control. The range of critical pressures for all refrigerants investigated extends from 380.8 psia for ethyl ether to 51,400 psia for mercury.

As regards the lower limit of evaporating temperature, the only temperature restriction obviates operational use below the freezing point for the fluid in question. However, pressure and specific-volume limitations become important as evaporator temperature decreases. Vapor cycles have not generally been operated below atmospheric pressure because of the difficulty in keeping air (a contaminant which adversely affects cycle performance) from leaking into the system. Toward the lower limit of evaporator temperature, the specific volume of the saturated vapor increases rapidly with decreasing temperature. For mercury, as an example, the value of this property increases from 8.4 to 5210 cu ft per lb as the evaporating temperature decreases from 1060 to 680 R. By necessitating enormous components, large fluid specific volumes possess especially little utility for aircraft refrigeration cycles.

Selection of Optimum Refrigerant in a Single Vapor Cycle. A plot of c.o.p. versus evaporator temperature with condenser temperature as a parameter may be prepared for each refrigerant used in actual vapor cycles. Fig. 5 plots the c.o.p. for Freon 11 with a compressor efficiency of 0.60.[5] The c.o.p. may be read from the figure for any combination of evaporator and condenser temperatures. In comparing c.o.p. when selecting optimum refrigerants for use over a particular temperature range, it is inconvenient first to prepare many graphs such as Fig. 5, one for each refrigerant, and then to read values of c.o.p. from them. A figure showing the relative c.o.p. for all refrigerants together gives a visual picture at a glance revealing which are thermodynamically best in each temperature range. In Fig. 6 actual c.o.p. for thirteen different refrigerants as well as the corresponding Carnot c.o.p have been plotted, each for cycles with condenser

[5] 0.60 was selected as a typical value for small rotary compressors.

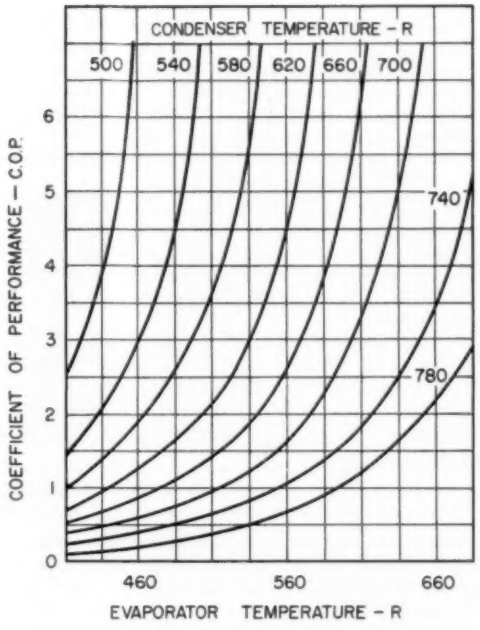


FIG. 5 FREON 11 VAPOR CYCLE PERFORMANCE

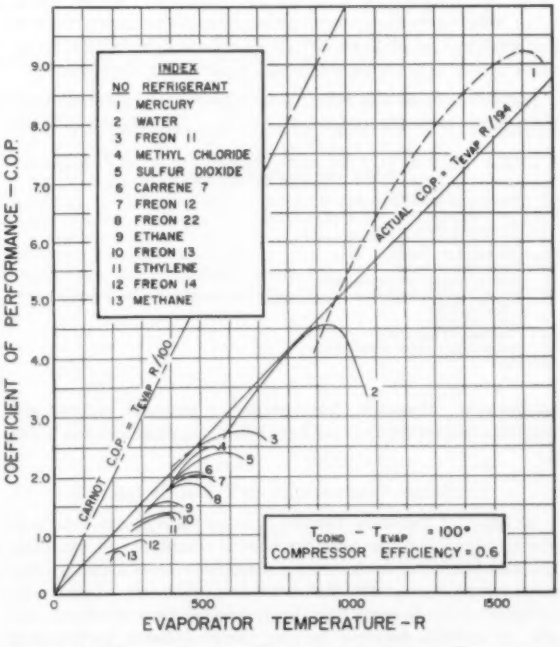


FIG. 6 C.O.P. FOR VARIOUS REFRIGERANTS VERSUS EVAPORATOR TEMPERATURE

temperatures 100 F higher than evaporator temperatures, and compresser efficiencies of 0.60. Inspection of Fig. 6 reveals some interesting effects:

1 Refrigerants employed in the same temperature range vary in performance for equal component efficiencies, allowing selection

of a thermodynamically optimum refrigerant for that range.

2 The highest values of the actual c.o.p. are consistently slightly greater than half of the values of Carnot c.o.p. throughout the entire temperature range considered. This would be expected with a compressor efficiency of 0.60 if the irreversibility of the expansion process causes little loss.

3 The working temperature range of each refrigerant appears to decrease with decreasing absolute temperature.

4 The c.o.p. of a 100 F cycle (of all cycles, in fact, independent of their evaporator-condenser temperature difference) exhibits a definite optimum point, falling off rapidly as the condenser temperature increases toward the refrigerant critical temperature, more gradually as the evaporator temperature decreases below the optimum point.

5 All c.o.p. tend to blend together on the same line as the difference between the refrigerant critical temperature and the cycle evaporator temperature increases for each refrigerant. However, in every case, this region for each refrigerant is accompanied by extremely small pressures and large vapor specific volumes, precluding use in this range.

The selection of a thermodynamically optimum refrigerant in a particular temperature range can be accomplished as follows:

(*a*) The range where refrigerant absolute pressure exceeds atmospheric pressure and refrigerant specific volume is lowest corresponds to the "hook" or the maximum c.o.p. part of each curve shown in Fig. 6.

(*b*) Selecting the appropriate temperature range on Fig. 6, observe which refrigerants exhibit the highest c.o.p. in this range. Select the refrigerant which exhibits the generally highest curve.

(*c*) Where two or more refrigerants appear extremely close in performance, calculations for each refrigerant in the particular cycle desired will reveal which exhibits better performance.

Additional Factors in Selecting Refrigerants. So far, only thermodynamic performance has been considered in selecting optimum refrigerants. However, the designer of practical systems will want to consider many other factors such as those listed in Table 2.

TABLE 2 CHECK LIST OF ADDITIONAL FACTORS FOR SELECTION OF REFRIGERANTS

1	Density of liquid and vapor
2	Viscosity of liquid and vapor
3	Thermal conductivity
4	Oil solubility
5	Chemical stability
6	Toxicity and odor
7	Reaction with water and oil
8	Reaction with materials used in cycle components and in the structure of the aircraft
9	Leak detection
10	Methods of handling
11	Other safety considerations
12	Cost

Series Combination of Vapor Cycles

Selection of Optimum Combinations of Refrigerants. As previously mentioned, the least mechanical complexity accompanies the least number of cycles used in series between two widely separated temperatures. Fig. 6 shows that a refrigerant selected with its optimum c.o.p. point near the desired condensing temperature will, in general, continue to give thermodynamic performance equivalent to other refrigerants at lower temperatures as its evaporator temperature is lowered indefinitely. In other words, if the critical temperature of a refrigerant exceeds the condensing temperature of the desired cycle, it will by itself give optimum thermodynamic performance, at the same time yielding least mechanical complexity. Unfortunately, however, decreasing saturation pressure and increasing vapor specific volume simultaneously dictate a practical lower limit of application for any refrigerant. In order to prevent operation at pressures less than atmospheric and at enormous values of specific volume requiring large apparatus volume and weight, it becomes necessary to construct multiple cycles, stacking them temperaturewise on top of each other as shown in Fig. 7.

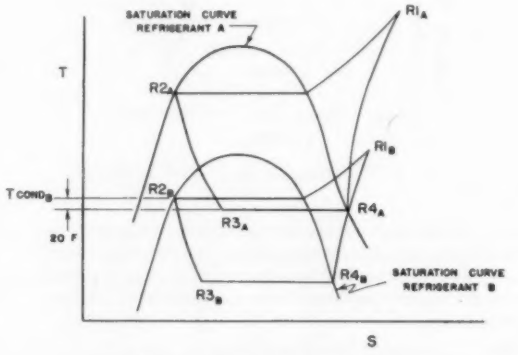


FIG. 7 BINARY VAPOR CYCLE

A series of refrigerants, each of optimum performance in its own temperature range, is selected to comprise a multiple cycle, remembering that in Fig. 6 the condenser temperature for each refrigerant is 100 deg F higher than the evaporator temperature at which the c.o.p. is plotted. As an example, in constructing an optimum multiple cycle between temperatures of 1000 and 200 R, one might choose mercury, Freon 11, Freon 13, and Freon 14 as an optimum combination. In general, use of these same refrigerants in widely spaced cycles requiring but two or three of them in series also would be justified, but the exact selection definitely depends on the end temperatures required. In single cycles, for instance, there are other refrigerants besides the four mentioned which work more successfully in the range 400–550 R with small evaporator-condenser temperature spreads because their particular optimum performance corresponds exactly to the temperature range considered.

Calculation of C.O.P. for Cycles in Series. The Carnot c.o.p. for a multiple cycle can be calculated by the same expression as given by Equation [2] since it is a function only of the end temperatures of the cycle, i.e., the highest condensing temperature and the lowest evaporating temperature. Besides the irreversibilities already present in the individual cycles, an additional irreversibility associated with the transfer of heat across the finite temperature difference between cycles appears. Because of the boiling and condensing processes taking place in these intermediate condenser-evaporators, the temperature difference will be relatively small, in the vicinity of 10–20 F. Using a temperature overlap of this order of magnitude, the multiple cycle c.o.p. can be calculated. The definition of c.o.p. for multiple vapor cycles becomes

$$\text{c.o.p.}_{\substack{\text{multiple}\\\text{vapor}\\\text{cycle}}} = \frac{Q_{\text{evaporator (lowest single cycle)}}}{\text{Total compressor work (all cycles in series)}} \quad [3]$$

For the binary cycle of Fig. 7 this expression becomes

$$\text{c.o.p.}_{\substack{\text{binary}\\\text{cycle}}} = \frac{Q_{\text{evap. (B)}}}{\text{compressor work } (a+b)} \quad [4]$$

$$= \frac{\text{c.o.p.}_a \times \text{c.o.p.}_b}{1 + \text{c.o.p.}_a + \text{c.o.p.}_b} \quad [5]$$

For a tertiary cycle, the expression becomes

$$\text{c.o.p.}_{\substack{\text{tertiary}\\\text{cycle}}} = \frac{Q_{(\text{lowest evap. temp.})(c)}}{\text{Total comp. work } (a + b + c)} \quad [6]$$

$$= \frac{\text{c.o.p.}_a \times \text{c.o.p.}_b \times \text{c.o.p.}_c}{\text{c.o.p.}_a \times \text{c.o.p.}_b + \text{c.o.p.}_b \times \text{c.o.p.}_c + \text{c.o.p.}_c \times \text{c.o.p.}_a + \text{c.o.p.}_a + \text{c.o.p.}_b + \text{c.o.p.}_c + 1} \quad [7]$$

Since the complexity of the multiple cycle c.o.p. equation obviously increases rapidly with the number of single cycles in combination, a more desirable procedure for obtaining numerical values of higher-order multiple-vapor cycle c.o.p. follows:

1 Determine the c.o.p. for each single cycle.

2 Start from one end of the series; combine the first two cycles, evaluating the c.o.p. of these two cycles with the binary-cycle equation.

3 Treating all previously considered cycles as a single cycle, continue to add single cycles using the binary-cycle equation repetitively.

The entire process, carried out for n-cycles in series, involves n calculations of single c.o.p. and $(n - 1)$ calculations of binary c.o.p.

Optimum Intermediate Evaporator-Condenser Temperatures. Fig. 6 demonstrates that the c.o.p. of virtually all refrigerants used in 100 F vapor cycles with constant enthalpy expansion, no line losses, and 60 per cent reversible adiabatic compression, tends to follow the straight line

$$\text{c.o.p.} = 5.15 \times 10^{-3} T_{\text{evap}} = \frac{T_{\text{evap}} R}{194} \quad [8]$$

or equivalently

$$\text{c.o.p.} = 0.515 \text{ (c.o.p) Carnot} \quad [9]$$

in a region distant from the critical point. Some refrigerants, however, appear to drop off in this range, so that their c.o.p. becomes less than that of the refrigerants in the next lower temperature range. The question then becomes: Can both of these representations be correct? No definite conclusion can be reached regarding this question, but since mercury is the chief offender in this respect, it is suspected that all refrigerants do tend toward the straight-line c.o.p. versus evaporator-temperature relationship in the region far from the critical point.[6] Two other refrigerants were found whose data showed extremely erratic performance in Fig. 6; both nitrous oxide and butane were omitted for this reason.[7] Where the data are suspected of inaccuracy, the lines have been shown dashed instead of solid in Fig. 6.

With the conclusion reached in the foregoing, there can be no optimum evaporator-condenser temperature between successive refrigerants in multiple cycles which results from considerations of thermodynamic performance alone. Instead, practical requirements limiting equipment size and pressure cause the designer to add another refrigerant to the series at or near the point where the saturation pressure of any refrigerant goes below atmospheric.

ACKNOWLEDGMENT

The authors gratefully acknowledge the assistance of the General Electric Company in supporting the program from which this paper is drawn.

REFERENCES

1 "Refrigeration Fundamentals," ASRE Data Book, seventh edition, 1951.

2 "Properties of Commonly-Used Refrigerants," ACRMA, 1948.

3 "Thermodynamic Properties of Steam," by J. H. Keenan and F. G. Keyes, John Wiley & Sons, Inc., New York, N. Y., 1936.

4 "Thermodynamic Properties of Mercury Vapor," by L. A. Sheldon, General Electric Technical Information Series Report DF 76054.

5 "Total Heat-Entropy Diagram for Mercury," by L. A. Sheldon, General Electric Company, April 15, 1948.

[6] The data for mercury are known not to be highly accurate, but probably the greater reason for the obvious discrepancy is the inaccurate large extrapolation made of the data.

[7] New revisions of data for some of the substances can be expected to improve the appearance of Fig. 6 in that a closer conformance to the straight line will probably be shown.

An Application of Complex Geometry to Relative Velocities and Accelerations in Mechanisms[1]

By G. H. MARTIN[2] and M. F. SPOTTS[3]

This paper describes an analytical method for determining velocities and accelerations in mechanisms by means of geometry of the complex plane. In addition to analysis of the four-bar linkage, a special solution is worked out for direct-contact mechanisms. The latter is included because, though accelerations in direct-contact mechanisms can be analyzed in terms of their equivalent four-bar linkages, there are certain cases where it is explained that the four-bar-linkage analysis breaks down. Further, it is shown that the method is general and is applicable to any mechanism having plane motion including complex mechanisms which cannot be treated as a series of four-bar linkages.

Introduction

USE of the complex plane for velocity and acceleration analysis in linkwork is not new. Beyer (1)[4] discusses an application of complex functions made by Bloch (2). Bloch developed a procedure for determining the relative lengths of the links for the four-bar linkage so that certain prescribed angular velocity and acceleration ratios between driving and driven crank would be satisfied. In a later paper, Rosenauer (3) has shown how the relative velocity and acceleration components occurring in various common linkages can be expressed in complex form.

The main purpose of this paper is to present a technique for handling in complex form the relative velocities and accelerations so that ultimately the velocities and accelerations of each of the links in a mechanism can be related to those of the driving member. The technique is not limited to the four-bar linkage, but applies also to complex linkages which cannot be analyzed as a series of four-bar linkages.

Though graphical solutions by the relative velocity and acceleration method are quicker than obtaining numerical results by the equations which are developed here, mathematical analysis nevertheless is advantageous under certain circumstances as, for example, when greater accuracy is desired. This would be particularly true for mechanisms in which the relative size of the links or instantaneous positions of the links is such that graphical results would be subject to large error because of the difficulty in accurately determining points of intersection on a drawing; for example, the point of intersection of nearly parallel lines. Further, mathematical analysis can be of value from the standpoint of systematic design.

The computation of the angular velocities and accelerations as well as the linear velocity and acceleration of any point on the linkage can be systematized and the substitution of numerical data into the equations developed herein can be carried out in tabular form. This would speed up the work if results were desired for a number of positions of the driving member. Solutions by tabulation can be carried out with the aid of a desk calculating machine by a person unskilled in kinematics once the form of the table is set up.

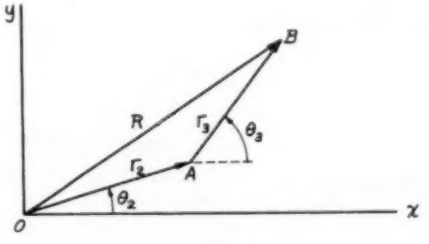


Fig. 1

In Fig. 1, points A and B are moving in the xy-plane relative to the fixed origin O. The position of A relative to O is specified by modulus r_2 and argument θ_2; the position of B relative to A is specified by means of modulus r_3 and argument θ_3. Vectors OA and AB are considered positive when directed as shown; the counterclockwise direction is positive for angles θ_2 and θ_3. In general r_2, r_3, θ_2, and θ_3 are all functions of time.

Resultant position vector R can be expressed as

$$R = OA + \rightarrow AB$$
$$= r_2 e^{i\theta_2} + r_3 e^{i\theta_3} \quad [1]$$

Should a third point C be present in Fig. 1, Equation [1] would be written with an additional term $r_4 e^{i\theta_4}$ to express the vector BC.

The exponentials in Equation [1] can be expanded and the real parts collected together and the imaginary parts collected together. Modulus and argument for R are then

$$\text{mod } R = \left\{ [\Sigma \mathcal{R}(R)]^2 + [\Sigma \mathcal{I}(R)]^2 \right\}^{1/2} \quad [2]$$

$$\arg R = \tan^{-1} \frac{\Sigma \mathcal{I}(R)}{\Sigma \mathcal{R}(R)} \quad [3]$$

Velocity vector $\dot{R}$ for point B is obtained by differentiation of Equation [1] with respect to time

$$\dot{R} = \dot{r}_2 e^{i\theta_2} + i r_2 \dot{\theta}_2 e^{i\theta_2} + \dot{r}_3 e^{i\theta_3} + i r_3 \dot{\theta}_3 e^{i\theta_3} \quad [4]$$

---

[1] From a dissertation submitted by the senior author in partial fulfillment of the requirements for the degree of Doctor of Philosophy, Northwestern University.

[2] Associate Professor of Mechanical Engineering, Michigan State University, East Lansing, Mich. Mem. ASME.

[3] Professor of Mechanical Engineering, Technological Institute, Northwestern University, Evanston, Ill. Mem. ASME.

[4] Numbers in parentheses refer to the Bibliography at the end of the paper.

Contributed by the Machine Design Division and presented at the Semi-Annual Meeting, Cleveland, Ohio, June 17–21, 1956, of The American Society of Mechanical Engineers.

Note: Statements and opinions advanced in papers are to be understood as individual expressions of their authors and not those of the Society. Manuscript received at ASME Headquarters, March 1, 1956. Paper No. 56—SA-32.

This equation also can be written in the following manner

$$\dot{R} = V_a^r +\!\!\rightarrow V_a^t +\!\!\rightarrow V_{b/a}^r +\!\!\rightarrow V_{b/a}^t \quad [4a]$$

where

V_a^r = absolute velocity of A in radial direction
V_a^t = absolute velocity of A in transverse direction
$V_{b/a}^r$ = velocity of B relative to A in radial direction
$V_{b/a}^t$ = velocity of B relative to A in transverse direction

These components, properly directed for positive quantities, are shown in Fig. 2. It should be noted in Equation [4] that a vector lies along the corresponding r except when the term is preceded by an i in which case the vector is rotated 90 deg counterclockwise from r.

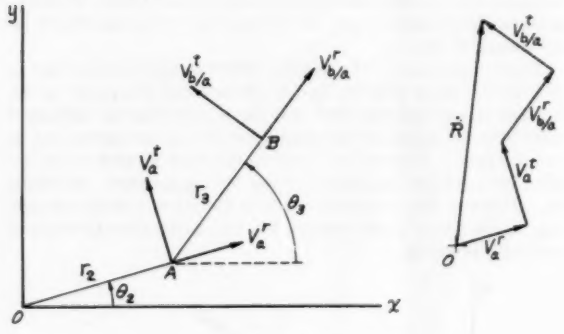


FIG. 2

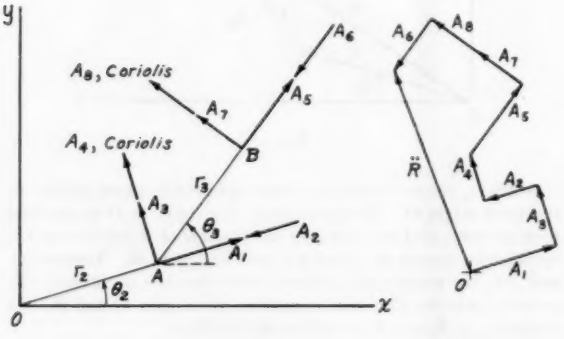


FIG. 3

We can expand Equation [4] in terms of its real and imaginary parts. The modulus and argument of $\dot{R}$ can then be expressed similar to Equations [2] and [3] except that R must be replaced by $\dot{R}$.

Acceleration vector $\ddot{R}$ for point B is found by differentiation of Equation [4] with respect to time

$$\ddot{R} = \ddot{r}_1 e^{i\theta_2} - r_2\dot{\theta}_2^2 e^{i\theta_2} + r_2\ddot{\theta}_2 i e^{i\theta_2} + 2\dot{r}_2\dot{\theta}_2 i e^{i\theta_2} + \ddot{r}_3 e^{i\theta_3} - r_3\dot{\theta}_3^2 e^{i\theta_3} + r_3\ddot{\theta}_3 i e^{i\theta_3} + 2\dot{r}_3\dot{\theta}_3 i e^{i\theta_3} \quad [5]$$

$$= A_1 +\!\!\rightarrow A_2 +\!\!\rightarrow A_3 +\!\!\rightarrow A_4 \text{ (Coriolis)} +\!\!\rightarrow A_5 +\!\!\rightarrow A_6 +\!\!\rightarrow A_7 +\!\!\rightarrow A_8 \text{ (Coriolis)} \quad [5a]$$

where

$A_1 +\!\!\rightarrow A_2 = A_a^r$ = absolute acceleration of A in radial direction
$A_3 +\!\!\rightarrow A_4 = A_a^t$ = absolute acceleration of A in transverse direction
$A_5 +\!\!\rightarrow A_6 = A_{b/a}^r$ = acceleration of B relative to A in radial direction
$A_7 +\!\!\rightarrow A_8 = A_{b/a}^t$ = acceleration of B relative to A in transverse direction

These eight components are represented in Fig. 3. It should be noted, because of the minus signs, that vectors A_2 and A_6 are directed in the opposite sense to vectors OA and AB.

From the real and imaginary parts of Equations [5] we can obtain expressions for the modulus and argument of $\ddot{R}$. They are similar to Equations [2] and [3] except that R is replaced by $\ddot{R}$.

Since many mechanisms can be reduced to an equivalent four-bar linkage or a combination of such linkages, the preceding equations are first applied to this basic type.

Four-Bar Linkage

In Fig. 4 lengths r_2, r_3, and r_4 as well as the angles θ_2, θ_3, and θ_4 are assumed to be known. The angular velocity $\dot{\theta}_2$ and angular

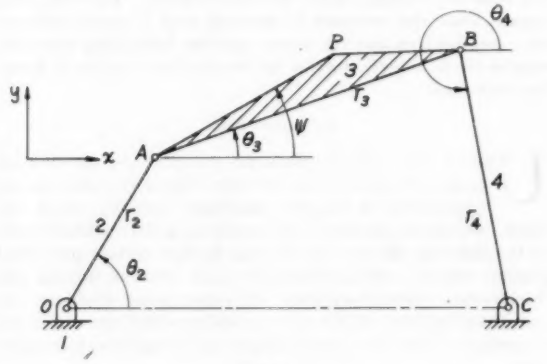


FIG. 4

acceleration $\ddot{\theta}_2$ are also known. Since r_2, r_3, and r_4 are constants, $\dot{r}_3 = \dot{r}_3 = \dot{r}_4 = 0$. Also, since the velocity of C relative to O is zero

$$V_{c/0} = V_{a/0} +\!\!\rightarrow V_{b/a} +\!\!\rightarrow V_{c/b} = 0$$

This may be written in the form of Equation [4] and can be expanded in terms of the real and imaginary parts as follows

$$\left.\begin{aligned} \Re(V_{c/0}) &= -r_2\dot{\theta}_2 \sin\theta_2 - r_3\dot{\theta}_3 \sin\theta_3 - r_4\dot{\theta}_4 \sin\theta_4 = 0 \\ \Im(V_{c/0}) &= r_2\dot{\theta}_2 \cos\theta_2 + r_3\dot{\theta}_3 \cos\theta_3 + r_4\dot{\theta}_4 \cos\theta_4 = 0 \end{aligned}\right\} \quad [6]$$

Equations [6] can be solved for the unknowns $\dot{\theta}_3$ and $\dot{\theta}_4$

$$\dot{\theta}_3 = -\frac{r_2 \sin\alpha}{r_3 \sin\beta}\,\dot{\theta}_2 \quad [7]$$

$$\dot{\theta}_4 = \frac{r_2 \sin\gamma}{r_4 \sin\beta}\,\dot{\theta}_2 \quad [8]$$

where $\alpha = \theta_2 - \theta_4$, $\beta = \theta_3 - \theta_4$, and $\gamma = \theta_2 - \theta_3$.

The velocity for any point P on link 3 may be written in the form of Equation [4]

$$V_p = V_{a/0} +\!\!\rightarrow V_{p/a} = ir_2\dot{\theta}_2 e^{i\theta_2} + ip\dot{\psi}e^{i\psi}$$

where

$$\rho = \overline{AP} \text{ and } \dot{\psi} = \dot{\theta}_3$$

$$\left.\begin{aligned} \mathcal{R}(V_p) &= -r_2\dot{\theta}_2 \sin\theta_2 - \rho\dot{\theta}_3 \sin\psi \\ \mathcal{I}(V_p) &= r_2\dot{\theta}_2 \cos\theta_2 + \rho\dot{\theta}_3 \cos\psi \end{aligned}\right\} \quad \ldots\ldots [9]$$

The magnitude and direction of V_p are then found from Equations [2] and [3] by substituting V_p for R.

The sum of the relative accelerations from O to C must equal zero. Thus

$$A_{c/o} = A_{a/o}{}^r +\!\!\rightarrow A_{a/o}{}^t +\!\!\rightarrow A_{b/a}{}^r +\!\!\rightarrow A_{b/a}{}^t +\!\!\rightarrow A_{c/b}{}^r +\!\!\rightarrow A_{c/b}{}^t = 0$$

This can be written in the form of Equation [5]. The terms containing $\dot{r}$ and $\ddot{r}$ quantities are equal to zero and the real and imaginary parts of the resulting equation become

$$\left.\begin{aligned} \mathcal{R}(A_{c/o}) &= -r_2\dot{\theta}_2{}^2 \cos\theta_2 - r_2\ddot{\theta}_2 \sin\theta_2 - r_3\dot{\theta}_3{}^2 \cos\theta_3 \\ &\quad - r_3\ddot{\theta}_3 \sin\theta_3 - r_4\dot{\theta}_4{}^2 \cos\theta_4 - r_4\ddot{\theta}_4 \sin\theta_4 = 0 \\ \mathcal{I}(A_{c/o}) &= -r_2\dot{\theta}_2{}^2 \sin\theta_2 + r_2\ddot{\theta}_2 \cos\theta_2 - r_3\dot{\theta}_3{}^2 \sin\theta_3 \\ &\quad + r_3\ddot{\theta}_3 \cos\theta_3 - r_4\dot{\theta}_4{}^2 \sin\theta_4 + r_4\ddot{\theta}_4 \cos\theta_4 = 0 \end{aligned}\right\} \quad \ldots [10]$$

Solving Equations [10] for $\ddot{\theta}_3$ and $\ddot{\theta}_4$, and substituting Equations [7] and [8], we have

$$\ddot{\theta}_3 = \frac{\dot{\theta}_2}{\dot{\theta}_3}\ddot{\theta}_2 - \frac{r_2\dot{\theta}_2{}^2 \cos\alpha + r_3\dot{\theta}_3{}^2 \cos\beta + r_4\dot{\theta}_4{}^2}{r_3 \sin\beta} \quad \ldots\ldots [11]$$

$$\ddot{\theta}_4 = \frac{\dot{\theta}_3}{\dot{\theta}_2}\ddot{\theta}_2 + \frac{r_2\dot{\theta}_2{}^2 \cos\gamma + r_3\dot{\theta}_3{}^2 + r_4\dot{\theta}_4{}^2 \cos\beta}{r_4 \sin\beta} \quad \ldots\ldots [12]$$

The linear acceleration of any point on the linkage may be found from a summation of the relative linear accelerations. For example, the acceleration of point P in Fig. 4 may be written in the form of Equation [5]

$$\left.\begin{aligned} A_p &= A_{a/o}{}^r +\!\!\rightarrow A_{a/o}{}^t +\!\!\rightarrow A_{p/a}{}^r +\!\!\rightarrow A_{p/a}{}^t \\ &= -r_2\dot{\theta}_2{}^2 e^{i\theta_2} + ir_2\ddot{\theta}_2 e^{i\theta_2} - \rho\dot{\theta}_3{}^2 e^{i\psi} + i\rho\ddot{\theta}_3 e^{i\psi} \\ \mathcal{R}(A_p) &= -r_2(\dot{\theta}_2{}^2 \cos\theta_2 + \ddot{\theta}_2 \sin\theta_2) - \rho(\dot{\theta}_3{}^2 \cos\psi \\ &\quad + \ddot{\theta}_3 \sin\psi) \\ \mathcal{I}(A_p) &= -r_2(\dot{\theta}_2{}^2 \sin\theta_2 - \ddot{\theta}_2 \cos\theta_2) - \rho(\dot{\theta}_3{}^2 \sin\psi \\ &\quad - \ddot{\theta}_3 \cos\psi) \end{aligned}\right\} \quad \ldots [13]$$

The magnitude and direction of A_p are then found from Equations [2] and [3] by substituting A_p for R.

The slider-crank mechanism in Fig. 5 is a special case of the four-bar linkage; i.e.

$$\theta_4 = \text{const} = \frac{3\pi}{2},\ r_4 = \infty, \text{ and thus } \dot{\theta}_4 = \ddot{\theta}_4 = 0$$

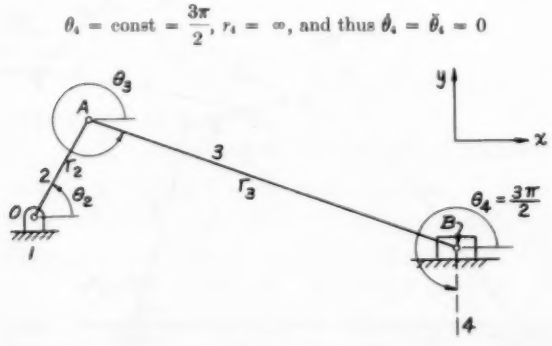


FIG. 5

Equations [7] and [11] give the angular velocity and acceleration respectively of the connecting rod.

There are certain direct-contact mechanisms in which one or two of the links in the equivalent four-bar linkage become infinite in length, and then the angular accelerations become indeterminate by Equations [11] and [12]. Examples are a disk cam with a flat-faced oscillating follower, and a disk cam having a reciprocating follower whose face is either curved or flat. Graphical analysis using the method of relative accelerations likewise is known to be inapplicable to the equivalent four-bar linkage in such cases. Graphical analysis is accomplished, however, by a consideration of the relative accelerations between the points in contact on cam and follower. The mathematical analysis presented in the following section likewise involves the relative accelerations for the contacting points, and the equations developed are applicable to any direct-contact mechanism.

Direct-Contact Mechanisms

Bodies 2 and 4 in Fig. 6 have arbitrary curves, and it is assumed that at all points their radii of curvature are known. A and B are points on 2 and 4, respectively, which are coincident at the instant considered. The lengths OA and BD together with θ_2, θ_4, and θ_T are assumed to be known. Either $\dot{\theta}_2$ and $\ddot{\theta}_2$ are known and $\dot{\theta}_4$ and $\ddot{\theta}_4$ are to be determined, or $\dot{\theta}_4$ and $\ddot{\theta}_4$ are known and $\dot{\theta}_2$ and $\ddot{\theta}_2$ are to be found.

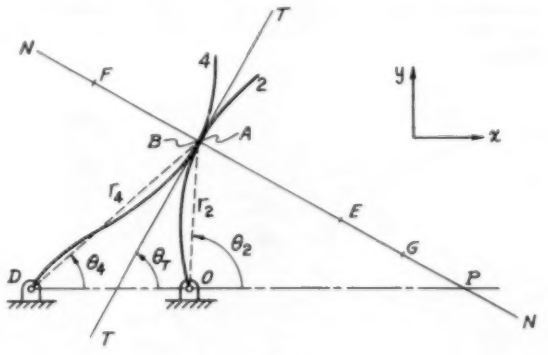


FIG. 6

Solution to the problem can be effected by considering point A as being temporarily fixed on link 2, with position vector AB of length $r_2 = 0$ lying along the tangent. Angle θ_T is measured from the x-axis and terminates with the position vector AB. The general layout is similar to that of Fig. 1. With these assumptions $\dot{r}_2 = \ddot{r}_2 = 0$, and Equation [4] as applied to Fig. 6 becomes

$$ir_4\dot{\theta}_4 e^{i\theta_4} = ir_2\dot{\theta}_2 e^{i\theta_2} + \dot{r}_2 e^{i\theta_T} \quad \ldots\ldots [14]$$

or

$$V_{b/d} = V_{a/o} +\!\!\rightarrow V_{b/a} \quad \ldots\ldots [15]$$

Expanding Equation [14] in terms of its real and imaginary parts

$$\left.\begin{aligned} -r_4\dot{\theta}_4 \sin\theta_4 + r_2\dot{\theta}_2 \sin\theta_2 - \dot{r}_2 \cos\theta_T = 0 \\ r_4\dot{\theta}_4 \cos\theta_4 - r_2\dot{\theta}_2 \cos\theta_2 - \dot{r}_2 \sin\theta_T = 0 \end{aligned}\right\} \quad \ldots [16]$$

Elimination of $\dot{r}_2$ from Equations [16] gives

$$\dot{\theta}_4 = \frac{r_2 \cos(\theta_2 - \theta_T)}{r_4 \cos(\theta_4 - \theta_T)}\dot{\theta}_2 \quad \ldots\ldots [17]$$

Next, elimination of $\dot{\theta}_4$ from Equations [16] gives

$$\dot{r}_3 = r_2\dot{\theta}_2 \frac{\sin(\theta_2 - \theta_4)}{\cos(\theta_T - \theta_4)} \quad \ldots\ldots\ldots\ldots [18]$$

Although the modulus of position vector AB in Fig. 6 is zero, the angular velocity of this vector relative to the tangent will in general not equal zero. Consider Fig. 7; r_{b2} is the radius of curvature at point A of the path which point B describes on body 2. Since A and B are coincident, angle θ_3 is equal to θ_T. However, B moves to B' in time interval dt with the chord length BB' equal to dr_3. The corresponding change in θ_3 is $d\theta_3$. Hence in Fig. 7

$$2d\theta_3 = \frac{dr_3}{r_{b2}}$$

This equation can be divided by dt and rearranged as

$$2\frac{d\theta_3}{dt} = \frac{1}{r_{b2}}\frac{dr_3}{dt}$$

or

$$\dot{\theta}_3 = \frac{1}{2r_{b2}}\dot{r}_3 \quad \ldots\ldots\ldots\ldots [19]$$

In Fig. 7, $dr_3 = \overline{BB'}$ is positive in sign and thus $\dot{r}_3 = (dr_3)/(dt)$ is positive. $d\theta_3$ is positive because it is measured counterclockwise from the tangent and thus $\dot{\theta}_3 = (d\theta_3)/(dt)$ is positive. Then for agreement in signs in Equation [19] r_{b2} must be considered positive when the center of curvature lies to the left of point B as shown.

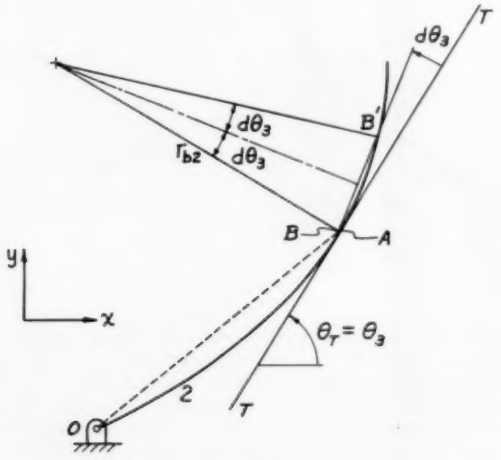


FIG. 7

The angular velocity of r_3 is then equal to $\dot{\theta}_2$ plus $\dot{\theta}_3$ with the latter determined by Equation [19]. Our general equation for acceleration, Equation [5], then becomes

$$-r_4\dot{\theta}_4^2 e^{i\theta_4} + r_4\ddot{\theta}_4 ie^{i\theta_4} = -r_2\dot{\theta}_2^2 e^{i\theta_2} + r_2\ddot{\theta}_2 ie^{i\theta_2}$$

$$+\ddot{r}_3 e^{i\theta_T} + 2\dot{r}_3\left(\dot{\theta}_2 + \frac{\dot{r}_3}{2r_{b2}}\right) ie^{i\theta_T} \quad \ldots\ldots [20]$$

$$A_{b/d^r} + \rightarrow A_{b/d^t} = A_2 + \rightarrow A_3 + \rightarrow A_4 + \rightarrow A_n$$

Expanding Equation [20] in terms of its real and imaginary parts, we obtain a pair of equations similar to Equations [16]. Solving them for $\ddot{\theta}_4$ and substituting the relationship expressed by Equation [17] we find

$$\ddot{\theta}_4 = \frac{\dot{\theta}_4}{\dot{\theta}_2}\ddot{\theta}_2 + \frac{1}{r_4\cos(\theta_4 - \theta_T)}\left[r_4\dot{\theta}_4^2\sin(\theta_4 - \theta_T) - r_2\dot{\theta}_2^2\sin(\theta_2 - \theta_T) + 2\dot{r}_3\left(\dot{\theta}_2 + \frac{\dot{r}_3}{2r_{b2}}\right)\right] \quad \ldots\ldots [21]$$

The Euler-Savary Equation may be used for determining r_{b2}. de Jonge (4) shows how this equation is easily derived using Hartmann's (5) construction for the center of curvature of the path of any point of a moving system. For Fig. 6, this equation becomes

$$\frac{1}{r'} + \frac{1}{r} = \frac{1}{r_0'} + \frac{1}{r_0} \quad \ldots\ldots\ldots\ldots [22]$$

where

r' = $\overline{BP}$ and P is the instant center for bodies 2 and 4. P lies at the intersection of the normal with line DO

r = $\overline{PG}$ and G is the center of curvature of the path which point B of body 4 describes on body 2

r_0' = $\overline{FP}$ where F is the center of curvature of body 4 at point B

r_0 = $\overline{PE}$ where E is the center of curvature of body 2 at point A

Then

$$r_{b2} = r' + r = \overline{BG} \quad \ldots\ldots\ldots\ldots [23]$$

Each of the quantities in Equations [22] and [23] is positive if it extends from 2 toward 4 along the normal.

Complex Mechanisms

The linkage in Fig. 8 is an example of a complex mechanism. The determination of the velocities or accelerations in a complex linkage consists of starting at some point of known velocity or acceleration, usually a point where these quantities are zero, and then traversing successive links summing up the relative velocity or acceleration components along the way in exponential form until another point of known velocity or acceleration is reached. The process is repeated until all the links in the mechanism have been traversed. As was done earlier, each vector equation is written in terms of its real and imaginary parts and thus provides two independent algebraic equations containing various unknown quantities. The foregoing procedure, in general, will provide a sufficient number of equations to permit a solution for the unknowns.

As an illustration of the procedure consider Fig. 8. It is as-

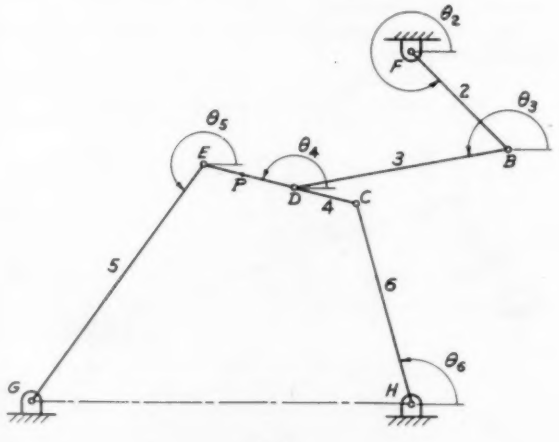


FIG. 8

sumed that θ_2, θ_3, θ_4, θ_5, θ_6, $\dot{\theta}_2$, and $\ddot{\theta}_2$ are known. The angular velocities and accelerations of all the links are to be found. Let

$r_2 = \overline{BF}$, $r_3 = \overline{BD}$, $r_4 = \overline{DE}$, $r_5 = \overline{EG}$, $r_6 = \overline{HC}$, and $r_7 = \overline{CE}$.

Velocities. Since points F, G, and H are points of zero velocity

$$V_{c/b} + V_{e/c} + V_{g/e} = 0$$

or

$$ir_6\dot{\theta}_6 e^{i\theta_6} + ir_7\dot{\theta}_4 e^{i\theta_4} + ir_5\dot{\theta}_5 e^{i\theta_5} = 0 \quad [24]$$

Also

$$V_{b/f} + V_{d/b} + V_{e/d} + V_{g/e} = 0$$

or

$$ir_2\dot{\theta}_2 e^{i\theta_2} + ir_3\dot{\theta}_3 e^{i\theta_3} + ir_4\dot{\theta}_4 e^{i\theta_4} + ir_5\dot{\theta}_5 e^{i\theta_5} = 0 \quad [25]$$

The sums of the real and imaginary parts of Equations [24] and [25] then give

$$-r_6\dot{\theta}_6 \sin\theta_6 - r_7\dot{\theta}_4 \sin\theta_4 - r_5\dot{\theta}_5 \sin\theta_5 = 0$$

$$r_6\dot{\theta}_6 \cos\theta_6 + r_7\dot{\theta}_4 \cos\theta_4 + r_5\dot{\theta}_5 \cos\theta_5 = 0$$

$$-r_2\dot{\theta}_2 \sin\theta_2 - r_3\dot{\theta}_3 \sin\theta_3 - r_4\dot{\theta}_4 \sin\theta_4 - r_5\dot{\theta}_5 \sin\theta_5 = 0$$

$$r_2\dot{\theta}_2 \cos\theta_2 + r_3\dot{\theta}_3 \cos\theta_3 + r_4\dot{\theta}_4 \cos\theta_4 + r_5\dot{\theta}_5 \cos\theta_5 = 0$$

After the numerical values of the known quantities are substituted into this set of equations, the latter can be solved for the values of $\dot{\theta}_3$, $\dot{\theta}_4$, $\dot{\theta}_5$, and $\dot{\theta}_6$.

Accelerations. Points F, G, and H have zero acceleration; thus

$$A_{c/b}{}^r + A_{c/b}{}^t + A_{e/c}{}^r + A_{e/c}{}^t + A_{g/e}{}^r + A_{g/e}{}^t = 0$$

or

$$-r_6\dot{\theta}_6{}^2 e^{i\theta_6} + r_6\ddot{\theta}_6 ie^{i\theta_6} - r_7\dot{\theta}_4{}^2 e^{i\theta_4} + r_7\ddot{\theta}_4 ie^{i\theta_4} - r_5\dot{\theta}_5{}^2 e^{i\theta_5} + r_5\ddot{\theta}_5 ie^{i\theta_5} = 0 \quad [26]$$

Also

$$A_{b/f}{}^r + A_{b/f}{}^t + A_{d/b}{}^r + A_{d/b}{}^t + A_{e/d}{}^r + A_{e/d}{}^t + A_{g/e}{}^r + A_{g/e}{}^t = 0$$

or

$$-r_2\dot{\theta}_2{}^2 e^{i\theta_2} + r_2\ddot{\theta}_2 ie^{i\theta_2} - r_3\dot{\theta}_3{}^2 e^{i\theta_3} + r_3\ddot{\theta}_3 ie^{i\theta_3} - r_4\dot{\theta}_4{}^2 e^{i\theta_4} + r_4\ddot{\theta}_4 ie^{i\theta_4} - r_5\dot{\theta}_5{}^2 e^{i\theta_5} + r_5\ddot{\theta}_5 ie^{i\theta_5} = 0 \quad [27]$$

Next, the sums of real and imaginary parts of Equations [26] and [27] are

$$-r_6\dot{\theta}_6{}^2 \cos\theta_6 - r_6\ddot{\theta}_6 \sin\theta_6 - r_7\dot{\theta}_4{}^2 \cos\theta_4 - r_7\ddot{\theta}_4 \sin\theta_4 - r_5\dot{\theta}_5{}^2 \cos\theta_5 - r_5\ddot{\theta}_5 \sin\theta_5 = 0$$

$$-r_6\dot{\theta}_6{}^2 \sin\theta_6 + r_6\ddot{\theta}_6 \cos\theta_6 - r_7\dot{\theta}_4{}^2 \sin\theta_4 + r_7\ddot{\theta}_4 \cos\theta_4 - r_5\dot{\theta}_5{}^2 \sin\theta_5 + r_5\ddot{\theta}_5 \cos\theta_5 = 0$$

$$-r_2\dot{\theta}_2{}^2 \cos\theta_2 - r_2\ddot{\theta}_2 \sin\theta_2 - r_3\dot{\theta}_3{}^2 \cos\theta_3 - r_3\ddot{\theta}_3 \sin\theta_3 - r_4\dot{\theta}_4{}^2 \cos\theta_4 - r_4\ddot{\theta}_4 \sin\theta_4 - r_5\dot{\theta}_5{}^2 \cos\theta_5 - r_5\ddot{\theta}_5 \sin\theta_5 = 0$$

$$-r_2\dot{\theta}_2{}^2 \sin\theta_2 + r_2\ddot{\theta}_2 \cos\theta_2 - r_3\dot{\theta}_3{}^2 \sin\theta_3 + r_3\ddot{\theta}_3 \cos\theta_3 - r_4\dot{\theta}_4{}^2 \sin\theta_4 + r_4\ddot{\theta}_4 \cos\theta_4 - r_5\dot{\theta}_5{}^2 \sin\theta_5 + r_5\ddot{\theta}_5 \cos\theta_5 = 0$$

The numerical values of all known quantities are then substituted into this set of equations. The latter can then be solved for the numerical values of $\ddot{\theta}_3$, $\ddot{\theta}_4$, $\ddot{\theta}_5$, and $\ddot{\theta}_6$.

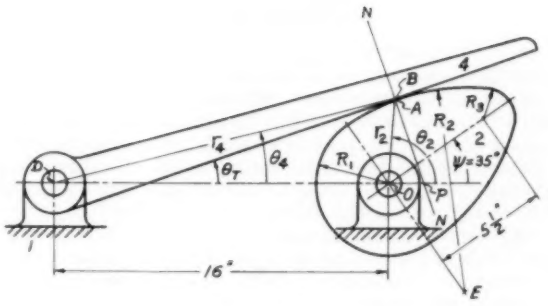


FIG. 9

Example. For the cam in Fig. 9, $R_1 = 3^1/_2$ in., $R_2 = 9^7/_8$ in., $R_3 = 1^1/_2$ in., $\dot{\theta}_2 = 3$ rad/sec, and $\ddot{\theta}_2 = -1.5$ rad/sec^2. The angular velocity and acceleration of link 4 are to be found for $\psi = 35$ deg.

Solution. From the trigonometry of the figure, $r_2 = 4.20$ in. $= 0.35$ ft; $r' = BP = 4.40$ in. $= 0.367$ ft; $r_4 = 17.0$ in. $= 1.42$ ft, $\theta_T = 19$ deg $= 0.332$ rad, $\theta_2 = 83.2$ deg $= 1.45$ rad, $\theta_4 = 14$ deg $= 0.244$ rad. From Equation [17]

$$\dot{\theta}_4 = \frac{0.35\cos(1.45 - 0.332)3}{1.42\cos(0.244 - 0.332)} = 0.326 \text{ rad/sec}$$

and from Equation [18]

$$\dot{r}_2 = 0.35(3)\frac{\sin(1.45 - 0.244)}{\cos(0.332 - 0.244} = 0.983 \text{ fps}$$

Substituting into Equation [22] we have

$$-\frac{1}{0.367} + \frac{1}{r} = \frac{1}{\infty} - \frac{1}{0.456}$$

or

$$r = 1.88 \text{ ft}$$

Then from Equation [23]

$$r_{b2} = -0.367 + 1.88 = 1.53 \text{ ft}$$

Equation [21] gives

$$\ddot{\theta}_4 = \frac{0.326}{3}(-1.5) + \frac{1}{1.42\cos(0.244 - 0.332)}\left[1.42(0.326)^2 \sin(0.244 - 0.332) - 0.35(9)\sin(1.45 - 0.332) + 2(0.983)\left(3 + \frac{0.983}{3.06}\right)\right]$$

$$= 2.46 \text{ rad/sec}^2$$

BIBLIOGRAPHY

1 "Kinematische Getriebesynthese," by R. Beyer, Springer-Verlag, Berlin, Germany, 1953, pp. 189–191.

2 "Syntheses of Four-Link Mechanisms," by Z. S. Bloch, *Izvestiya Academiyi Nauk USSR Otdeleniye Technicheskikh Nauk* 1940, No. 1, pp. 47–54.

3 "Gelenkvierecke mit vorgeshriebenen Groesst- und Kleinstwerten der Abtriebswinkelgeschwindigkeit," by N. Rosenauer, *Werkstattstechnik*, vol. 38, 1944, pp. 25–27.

4 "A Brief Account of Modern Kinematics," by A. E. R. de Jonge, Trans. ASME, vol. 65, 1943, pp. 663–683.

5 "Ein Neues Verfahren zur Aufsuchung des Krummungkreises," by W. Hartmann, *Zeitschrift des Vereines deutscher Ingenieure*, vol. 37, 1893, pp. 95–101.

Discussion

WM. J. CARTER.[5] Velocity and acceleration analysis of ordinary four-bar mechanisms by the complex number vector method is not new, as the authors state. The writer has used this method for velocity and acceleration analysis and also for the determination of higher time derivatives of motion. The writer shares the author's opinions about the advantages and disadvantages of this method. The wider usage of digital computing equipment should make this method more attractive.

The authors are to be commended for the extension of the complex number vector method to the analysis of direct contact mechanisms. This feature is new so far as the writer is aware.

It is of interest to note that the author's Equation [17] may be established from the known properties of the pitch point P. Let $\overline{DK}$ and $\overline{OH}$ be the lengths of the perpendiculars from D and O respectively to the line N-N. It may be seen that

$$\frac{\overline{OH}}{\overline{DK}} = \frac{r_2 \cos(\theta_2 - \theta_T)}{r_4 \cos(\theta_4 - \theta_T)} = \frac{\overline{OP}}{\overline{DP}} = \frac{\dot{\theta}_4}{\dot{\theta}_2} \quad \ldots\ldots [28]$$

and

$$\dot{\theta}_4 = \dot{\theta}_2 \frac{\overline{OP}}{\overline{DP}} = \frac{r_2 \cos(\theta_2 - \theta_T)}{r_4 \cos(\theta_4 - \theta_T)} \dot{\theta}_2 \quad \ldots\ldots [29]$$

which is the same as the author's Equation [17]. One might now obtain an expression for $\ddot{\theta}_4$ by differentiation of Equation [17] with respect to time. This would, of course, lead to the same result as that given by the authors, although the physical interpretation would now involve the time derivatives of $\overline{DP}$ and $\overline{OP}$. Koenig[6] has used this method for the analysis of the ordinary four-bar linkage.

Allen H. Candee, in his discussion of a recent paper by Freudenstein[7] has analyzed the direct contact problem by a method which is slightly different from that of the authors or the method of Koenig.

It should be noted that the vector, $r_3 = 0$ in Fig. 6, may not be considered as forming an equivalent four-bar linkage with r_2 and r_4 in the normal sense. If such were the case the crossing of T-T and $\overline{DO}$ would be the pitch point which is obviously incorrect.

FERDINAND FREUDENSTEIN.[8] The introduction of complex variable techniques in kinematics in 1940 by S. Sh. Blokh came about as a result of investigations in kinematic synthesis, which field constitutes the major portion of modern kinematics. Up to the present there has been no major effort involving complex variables in mechanisms synthesis work in this country, although the potentialities of the method are great. To the extent that the authors, while concerned with kinematic analysis, focus attention on the use of complex variables, they have performed a service to the profession.

The basic development (i.e., the Introduction) represents no new ideas or principles as is evident from a perusal of pp. 189–195 of R. Beyer's "Kinematische Getriebesynthese" (authors' ref. 1)—as the authors well know—but the derivations using complex variables, which follow, have not previously appeared in Western European or American technical literature as far as is known to this discusser. Some of the velocity equations can be read off at sight by considering the equivalent four-bar linkage and using the well-known theorem relating the angular velocity ratio to the ratio of the perpendiculars drawn from the cranks to intersect the connecting link. Considerable relevant work along the same lines can be found in the writings of S. Sh. Blokh and these have been reviewed in the *Applied Mechanics Reviews*. Some specific comments follow:

(*a*) The description of references (1) and (3) leaves room for improvement. Beyer (1) describes an application of complex variables to the synthesis of four-bar linkages having prescribed values of the angular velocities and angular accelerations of the links. Rosenauer (3) describes an extension of this technique to the synthesis of linkages having prescribed extreme values of the angular velocity ratio of driving and driven links.

(*b*) In using the equations developed by the authors, it may be found desirable to summarize the equations in easy-to-use tabular form. This would permit the use of modern computational facilities.

(*c*) A method used by S. Sh. Blokh and more recently also by K. H. Sieker utilizes complex conjugates to eliminate unwanted unknowns. In a four-bar linkage, for instance, usually only the displacement, θ_2, of the driving crank and the lengths of the links are known; in deriving the angular velocity and acceleration of the driven link it would be convenient, therefore, if no quantities other than these were to enter the equations. This can be done as follows:

Let $r_1 = OC$, Fig. 4. Chain closure is expressed, therefore, by the equation

$$r_2 e^{i\theta_2} + r_3 e^{i\theta_3} + r_4 e^{i\theta_4} = r_1$$

We can write in addition the complex conjugate form of this equation which applies to the mechanism reflected about the fixed link, OC

$$r_2 e^{-i\theta_2} + r_3 e^{-i\theta_3} + r_4 e^{-i\theta_4} = r_1$$

We can eliminate "unwanted" θ_3 from these equations, obtaining

$$\left(\frac{r_1^2 + r_2^2 - r_3^2 + r_4^2}{r_2 r_4}\right) - \left(\frac{r_1}{r_4}\right)$$

$$(e^{i\theta_2} + e^{-i\theta_2}) - \frac{r_1}{r_2}(e^{i\theta_4} + e^{-i\theta_4}) + (e^{i(\theta_2-\theta_4)} + e^{-i(\theta_2-\theta_4)}) = 0$$

Upon differentiation of this equation, it can be shown that

$$\dot{\theta}_4 = \dot{\theta}_2 \frac{r_2 r_4 \sin(\theta_2 - \theta_4) - r_1 \sin\theta_2}{r_4 r_2 \sin(\theta_2 - \theta_4) + r_1 \sin\theta_4}$$

$$\ddot{\theta}_4 = \frac{\ddot{\theta}_2 \dot{\theta}_4}{\dot{\theta}_2} - \frac{r_1 r_2 \dot{\theta}_2^2 \cos\theta_2 + r_1 r_4 \dot{\theta}_4^2 \cos\theta_4 - r_2 r_4 (\dot{\theta}_2 - \dot{\theta}_4)^2 \cos(\theta_2 - \theta_4)}{r_2 r_4 \sin(\theta_2 - \theta_4) + r_1 r_4 \sin\theta_4}$$

of course these equations still include θ_4.

These equations should be compared with Equations [8] and [12] of the authors. This discusser respectfully suggests that the authors consider the possibilities of this technique in their future work, especially in the field of kinematic synthesis.

A. S. HALL, JR.[9] The increasing availability of digital computing equipment makes it profitable to re-examine our ways of doing things in many fields of engineering. We have barely touched upon the possibilities for using such equipment in kinematic analysis and synthesis. The authors have performed

[5] Associate Professor, Department of Mechanical Engineering, The University of Texas, Austin, Texas. Mem. ASME.

[6] "A Uniform Method for Determining Angular Accelerations," by L. R. Koenig, *Journal of Applied Mechanics*, Trans. ASME, vol. 68, 1946, pp. A-41–44.

[7] "On the Maximum and Minimum Velocities and the Accelerations in Four-Link Mechanisms," by F. Freudenstein, Trans. ASME, vol. 78, 1956, pp. 779–787.

[8] Assistant Professor, Department of Mechanical Engineering, Columbia University, New York, N. Y. Assoc. Mem. ASME.

[9] Professor, School of Mechanical Engineering, Purdue University, Lafayette, Ind. Mem. ASME.

a service in spelling out an approach to writing velocity and acceleration equations which can easily be programmed for automatic computation. If we are to go in for automatic computation of velocities and accelerations, then we may as well include *position*. The authors have assumed position data known in all their example problems, perhaps feeling this problem to be trivial.

In connection with the example on direct contact mechanisms perhaps a word might be added on the subject of "equivalent linkages." It is implied in the paper (as well as in most textbooks) that there is *one* equivalent linkage. Actually there exists an infinite number of four-bar linkages equivalent, insofar as instantaneous velocities and accelerations are concerned, to a given direct contact mechanism. To form such an equivalent four-bar we could choose, quite arbitrarily, a point M on body 2 (see Fig. 6 of the paper) to be the location of the pin joint between 2 and the connecting-rod of the equivalent four-bar. Then, using the Euler-Savary equation, we could solve for the location N of the center of curvature of the path which M_2 traces in the motion of 2 relative to 4. If we then let N be the location of the pin joint between 4 and the connecting-rod we shall have our equivalent mechanism.

The authors have actually found one such equivalent mechanism in their example problem, Fig. 6. It is formed by pinning a connecting-rod between B on body 4 and G on body 2.

This somewhat larger view of the equivalent linkage does make it possible to reduce the direct contact problem to the four-bar problem, even for the case in which one of the direct contact bodies has infinite radius of curvature.

R. T. Hinkle.[10] The authors state that the theory presented in this paper for the determination of velocities and accelerations is not new. However, they have generalized it by extending it to include direct contact and complex mechanisms. Since other analytical methods for determining acceleration become so cumbersome when applied to complex mechanisms as to be almost useless, this paper is a major contribution.

The graphical method will probably remain in wide use for velocity and acceleration analysis, but when synthesis is considered, some form of analytical approach is needed. It may be that this paper will have its greatest value in the development of this field. At the present time, one investigator is applying the theory presented here for the determination of inertia stresses in mechanisms.

Authors' Closure

Professor Carter points out that Equation [17] which applies to direct contact mechanisms may also be established from the known properties of the pitch point P in Fig. 6. Further, he states that this equation may be differentiated to obtain an expression for $\ddot{\theta}_4$, a method which Koenig has used in an analysis of the ordinary four-bar linkage. This is true. However, if we were to follow this suggestion we would find that the resulting expression obtained for $\ddot{\theta}_4$ would contain the quantities $(\dot{\theta}_2 - \dot{\theta}_T)$ and $(\dot{\theta}_4 - \dot{\theta}_T)$. The values of both of these quantities are unknown of course because $\dot{\theta}_T$ is not known. Thus it appears that such a procedure would not provide a solution for $\ddot{\theta}_4$.

In commenting on $r_2 = 0$ in Fig. 6, Professor Carter warns that this is not to be regarded as link 3 in an equivalent four-bar linkage. True, the authors have not used r_3 as such and appreciate that this discusser's comment should help further in preventing any misconception of this kind by the reader.

Professor Freudenstein states that some of the velocity equations can be read off at sight by considering the equivalent four-bar linkage and using the well-known theorem relating the angular velocity ratio to the ratio of the perpendiculars drawn from the cranks to intersect the connecting link. This is the same comment as is made by Professor Carter. The authors have been aware of this method, but throughout the paper the approach to both velocity and acceleration analysis has been by means of summation of relative velocity and acceleration vectors. This has been done to demonstrate the application of the general expressions presented in the Introduction.

The use of complex conjugates as pointed out by Professor Freudenstein for eliminating unknown θ_3 in the four-bar linkage is indeed a further gain even though his equations for $\dot{\theta}_4$ and $\ddot{\theta}_4$ contain more terms than the authors' Equations [8] and [12].

The comments by Professor Hall as to the desirability of expressing position in a form which would be convenient for computational purposes, as the authors have done for velocity and acceleration, are well placed. The authors' equations give the velocity and acceleration of any link in the mechanism in terms of the velocity and acceleration of the driving link alone, and it has been shown how the unknown velocities and accelerations of the other links in the kinematic chain can be eliminated. However, using exponentials, no means has been found to eliminate a sufficient number of unknown angles so that the position of any one link can be expressed in terms of the lengths of the links and the angular position of the driver alone. The method of using complex conjugates, as cited by Professor Freudenstein in his discussion, eliminates "unwanted" θ_3 from the position vector equation for the four-bar linkage. But, as he mentions, unwanted θ_4 still remains. The method commonly used in finding the angular positions of the links has been to use ordinary trigonometry. We know for example that in relating angles and lengths in the four-bar linkage it has been customary to work with the diagonal on the linkage. This is the method the authors have used after devoting considerable effort to find more efficient methods.

The comment by Professor Hall that there is an infinite number of equivalent four-bar linkages for a direct-contact mechanism should be emphasized indeed. Professor Hall has explained how we can obtain equivalent four-bar linkages where all links are of finite length even for cams having flat-faced followers. Thus in addition to the solution of velocities and accelerations for direct-contact mechanisms as expressed by Equations [17] and [21], any direct-contact mechanism may be analyzed using the four-bar-linkage solution as expressed by Equations [7], [8], and [12].

In conclusion, the authors wish to thank Professors Carter, Freudenstein, Hall, and Hinkle for their contributions and for the trouble they have taken in discussing the paper.

[10] Professor of Mechanical Engineering, Michigan State University, East Lansing, Mich. Mem. ASME.